The Finite-Time Exergy Limit of Computation: Reconciling Landauer's Principle with the Speed-Efficiency Paradox

Grigory G. Potapov[1]
D. Mendeleev University of Chemical Technology of Russia (MUCTR)
August 2026

Abstract

Modern computing hardware operates far from thermodynamic equilibrium, yet the fundamental energy limits of computation are still assessed using the quasistatic Landauer bound, which assumes infinitely slow, reversible operations. Here, we develop a finite-time exergy framework that unifies Landauer's principle with the Gouy–Stodola theorem and linear non-equilibrium thermodynamics to quantify the continuous generation of anergy—the irreversibly degraded fraction of energy—in realistic computing systems. By partitioning the hardware topology into three thermodynamic zones and applying the Onsager formalism to the microscopic substrate, we derive a master equation that decouples the informational, dynamic, and static contributions to computational dissipation as explicit functions of the finite operation time τ. We prove that the competition between dynamic transport friction ($\propto 1/\tau^2$) and static leakage ($\propto \tau$) produces a unique thermodynamic optimum τ^*_{opt}, and that the non-linear thermal-leakage feedback generates a saddle-node bifurcation defining the absolute physical speed limit of any dissipative substrate. Applying the framework to a projected 2026 sub-2 nm 3D-IC tensor core and analog ReRAM crossbars, we show that for this specific architecture, the optimal clock frequency converges to 2.36 GHz under realistic thermal constraints. Furthermore, its absolute bifurcation limit at 16.1 GHz is fundamentally coupled to the Landauer informational cost. Our results establish that post-Moore scaling must prioritise macroscopic exergy minimisation over raw frequency scaling.

Keywords: Finite-time thermodynamics, Landauer's principle, Exergy, 3D-IC, Neuromorphic computing, Thermal runaway, Speed-efficiency paradox.

1. Introduction

The relentless scaling of semiconductor technology has driven computing performance to unprecedented heights, yet it has simultaneously pushed physical substrates to the brink of their macroscopic thermodynamic limits. While Landauer's principle establishes a fundamental lower bound on the energy required for logical erasure (Landauer, 1961; Berut et al., 2012; Chattopadhyay et al., 2025), modern high-performance architectures—ranging from sub-2 nm three-dimensional integrated circuits (3D-ICs) to analog neuromorphic crossbars—operate far from this quasistatic ideal (Wong & Salahuddin, 2015). In these highly non-equilibrium regimes, the dynamic cost of spatial data transport and the static penalty of thermal leakage dominate the energy landscape, creating a severe speed-efficiency paradox. Recent projections indicate that the

[1] e-mail: potapovgri@gmail.com
ORCID: 0009-0008-5543-1816

energy demands of large-scale AI inference are approaching macroscopic limits, threatening global energy infrastructure (Oviedo et al., 2025; Tkachenko, 2025; IEA, 2026; Yadav et al., 2026). To sustain post-Moore scaling, it is no longer sufficient to analyze logic gates in isolation; we must rigorously quantify the continuous generation of irreversibility across the entire hardware-cooling hierarchy. Here, we introduce a finite-time computational exergy framework that bridges microscopic stochastic thermodynamics with macroscopic heat rejection, revealing the absolute physical speed limits of dissipative computing.

2. Theoretical Framework: The Generalized Exergy Balance of Finite-Time Computation

To bridge the gap between theoretical physics and practical hardware engineering, we must formalize the physical nature of information processing, spatial data transport, and energy degradation. According to the First Law of Thermodynamics, total energy U is always conserved. However, the Second Law dictates that the quality of this energy continuously degrades during any real process. To capture this, we partition the total internal energy into two fundamental components (Rant, 1956; Moran et al., 2014; Dincer & Rosen, 2020): Exergy (E), representing the maximum useful work potential of a system relative to the environmental dead state, and Anergy (A), representing the fraction of energy irreversibly degraded into waste heat.

While energy is conserved, exergy is continuously destroyed by physical irreversibilities. According to the Gouy-Stodola theorem (Gouy, 1889; Stodola, 1904; Bejan, 1996), the continuous generation of anergy is mathematically defined as:

$$dA = T_0 dS_{gen} \qquad (2.1)$$

where dA is the anergy generation rate, T_0 is the absolute temperature of the ambient environment (the dead state), and dS_{gen} is the total entropy generation rate. In modern stochastic thermodynamics (Peliti & Pigolotti, 2021; Wolpert et al., 2024; Falasco & Esposito, 2025), information acts as a physical constraint on the accessible phase space of the computing substrate. When a logical state is irreversibly erased, the informational exergy is annihilated. Consequently, the fundamental thermodynamic penalty of computation is not merely the 'heat' generated within the processor, but the discrete generation of anergy paid directly to the environmental sink at T_0 (Cobos-Murcia, 2026), unless reversible computing paradigms are employed (DeBenedictis et al., 2017; Frank, 2017; DeBenedictis, 2025).

2.1. The Generalized Exergy Balance and the Three-Zone Thermal Hierarchy

Modern computing hardware infrastructure is a multi-scale thermodynamic system. To rigorously track the continuous generation of anergy and identify physical bottlenecks, we partition the

hardware topology into three distinct thermodynamic zones. Crucially, we must distinguish between the physical origins of energy dissipation, the mechanisms of inter-zone heat transfer, and the dual nature of environmental heat exchange (natural/useful vs. parasitic).

Zone 1: The Microscopic Substrate (The Chip).

This zone represents the active silicon or quantum substrate. The input to Zone 1 is the total electrical power $\dot{W}_{in}$ drawn from the power delivery network, which is 100% pure exergy rate. Physically, this electrical power is bifurcated into two distinct processes before it entirely thermalizes:

1. **Electrical Delivery and Spatial Transport ($\dot{W}_{elec}$):** The energy consumed by the physical routing of data across interconnects, overcoming parasitic capacitances, and the Joule heating associated with current flow through the transistor channels and metal lines.
2. **Logical Computation ($\dot{W}_{comp}$):** The energy strictly associated with logical state transitions, phase-space compression, and the fundamental informational erasure of bits.

Thus, $\dot{W}_{in} = \dot{W}_{elec} + \dot{W}_{comp}$. Both components ultimately degrade into heat, but $\dot{W}_{comp}$ carries the fundamental informational anergy penalty, while $\dot{W}_{elec}$ represents the macroscopic electro-spatial friction (Das et al., 2024; Virmani & Chatterjee, 2025).

This zone is characterized by the effective exergetic junction temperature, T_{eff}. While standard volumetric or arithmetic temperature averages are computationally convenient, they fundamentally violate the Second Law of Thermodynamics when applied to modern 3D-ICs characterized by extreme spatial non-uniformity (Moran et al., 2014). To ensure strict thermodynamic fidelity bridging nanoscale stochastic heat generation with macroscopic heat rejection, T_{eff} must be rigorously defined as the entropy-flux-preserving effective temperature (Callen, 1985; Seifert, 2025). This formulation mathematically anchors T_{eff} to the bulk active area, preventing negligible extreme hotspots from artificially skewing the macroscopic exergy balance.

$$\frac{P_{total}}{T_{eff}} = \int_V \frac{p(\boldsymbol{r})}{T(\boldsymbol{r})} dV \qquad (2.2)$$

where $p(\boldsymbol{r})$ and $T(\boldsymbol{r})$ are the local power density (in W/m^3) and local temperature, respectively, and P_{total} is the total dissipated power (see Supplementary Note S2.1 for the explicit volumetric integral derivation). This formulation mathematically 'anchors' T_{eff} to the bulk active area where

the vast majority of the power is dissipated, ensuring strict fidelity when bridging nanoscale non-equilibrium phonon dynamics with macroscopic heat rejection.

While Equation 2.2 relies on the classical assumption of local thermal equilibrium, in sub-2nm CFET architectures (Seshanth et al., 2026), the characteristic length scales shrink below the phonon mean free path. This triggers quasi-ballistic, non-Fourier heat transport, where phonons travel from the hotspot to the boundaries without thermalizing, significantly reducing the effective thermal conductivity (Liu et al., 2024; Seshanth et al., 2026; Kim et al., 2026). This transition from diffusive to ballistic phonon transport fundamentally alters energy dissipation at the nanoscale, a phenomenon extensively documented in modern semiconductor physics (Pop, 2010; Moore & Shi, 2014). In this regime, defining a local thermodynamic temperature becomes physically ambiguous. To resolve this, our framework treats T_{eff} not as a local microscopic temperature, but as a macroscopic, lumped exergy-accounting parameter mathematically anchored to preserve the total entropy flux at the package boundary.

Thermal (Energy) Balance for Zone 1:

Assuming steady-state conditions and no mechanical work output, the First Law of Thermodynamics dictates that the input electrical power $\dot{W}_{in}$ must be exactly balanced by the net heat transfer rates crossing the control volume boundary (Bejan, 1996; Moran et al., 2014). In a realistic, non-adiabatic package, the chip interacts with both the dedicated cooling infrastructure (Zone 2) and the universal ambient environment (the dead state at T_0). The generalized energy balance is:

$$\dot{W}_{in} = \dot{Q}_{1\to 2} + \dot{Q}_{env,1} \qquad (2.3)$$

where $\dot{Q}_{1\to 2}$ is the primary heat flux directed to the cooling system, and $\dot{Q}_{env,1}$ represents the direct unmediated heat exchange with the ambient environment (e.g., thermal conduction through the PCB or natural convection from the package periphery).

This generalized balance proves that environmental interaction is not merely a negligible artifact, but a fundamental thermodynamic vector that either mitigates (in classical systems) or severely compounds (in quantum systems) the macroscopic cooling penalty.

Exergy Balance for Zone 1:

The local exergy balance within this microscopic control volume must strictly reflect the bifurcation of the thermal fluxes established in Equation 2.3. The input electrical power $\dot{W}_{in}$ (which is 100% pure exergy rate) is partitioned into three components: the irreversibly generated Anergy within the chip ($\dot{A}_{gen,1}$), the residual thermal exergy of the primary heat stream directed

to the coolant ($\dot{E}_{1\to2}^{(Q)}$), and the exergy completely destroyed by the unmediated heat exchange with the environment ($\dot{A}_{loss,1}$):

$$\dot{W}_{in} = \dot{A}_{gen,1} + \dot{E}_{1\to2}^{(Q)} + \dot{A}_{loss,1} \qquad (2.4)$$

The residual exergy of the primary heat stream is governed by a universal positive Carnot penalty function evaluated at T_{eff}:

$$\dot{E}_{1\to2}^{(Q)} = \dot{Q}_{1\to2}\left(1 - \frac{T_0}{T_{eff}}\right) \qquad (2.5)$$

Similarly, the thermal energy $\dot{Q}_{env,1}$ leaving the chip carries an initial exergy potential of $\dot{Q}_{env,1}\left(1 - T_0/T_{eff}\right)$. However, because this unmediated flux bypasses the cooling infrastructure and thermalizes directly with the dead state at T_0, its entire work potential is annihilated. Thus, this parasitic exergy loss is strictly defined as $\dot{A}_{loss,1} = \dot{Q}_{env,1}\left(1 - T_0/T_{eff}\right)$.

This explicit formulation demonstrates a crucial thermodynamic principle: T_{eff} strictly dictates the proportion of the input electrical power that is irreversibly destroyed within the silicon versus the thermodynamic potential that survives to reach the primary packaging boundary.

This generalized exergy formulation inherently captures both classical and cryogenic regimes, as the Carnot factor $\left(1 - T_0/T_{eff}\right)$ mathematically adapts to the direction of the thermal gradient:

- **In the classical regime ($T_{eff} > T_0$):** The chip is hotter than the ambient environment, meaning parasitic heat naturally dissipates outward ($\dot{Q}_{env,1} > 0$). Concurrently, the Carnot factor $\left(1 - T_0/T_{eff}\right)$ is positive. Consequently, $\dot{E}_{1\to2}^{(Q)}$ represents a positive residual work potential of the rejected heat stream. The unmediated outward heat flux acts as a beneficial secondary cooling path, but its thermalization with the ambient environment results in a strictly positive exergy loss $\dot{A}_{loss,1}$.
- **In the cryogenic regime ($T_{eff} < T_0$):** The chip operates at deep-cryogenic temperatures, causing ambient heat to leak into the substrate ($\dot{Q}_{env,1} < 0$). In this regime, the Carnot factor becomes strictly negative. Consequently, $\dot{E}_{1\to2}^{(Q)}$ emerges as a negative quantity, representing an exergy deficit—macroscopic work must be supplied by the cryoplant to pump this heat. Furthermore, the product of the negative heat flux and the negative Carnot factor yields a strictly positive $\dot{A}_{loss,1}$ rigorously quantifying the massive anergy generated when room-temperature heat irreversibly thermalizes with the cryogenic substrate (Zhang et al., 2024; Dutta et al., 2025; Trupiano et al., 2026).

Zone 2: The Mesoscopic Thermal Management System (The Coolant)

This zone encompasses the primary physical cooling infrastructure directly interfacing with the chip (e.g., air heat sinks with local fans, liquid cooling loops, two-phase microfluidic channels (Kanbur et al., 2021), or the primary stages of cryogenic refrigerators). Unlike the solid-state substrate of Zone 1, Zone 2 operates as an open thermodynamic flow system where a fluid (air, liquid, or cryogenic mixture) acts as the heat carrier.

Energy Balance for Zone 2:

The inputs to Zone 2 include the primary heat flux from the chip ($\dot{Q}_{1\to2}$), the enthalpy of the incoming coolant ($\dot{H}_{2,in}$), and the macroscopic infrastructure power ($\dot{W}_{infra,2}$) supplied to active components (McGovern, 2015). In classical air-cooled systems, $\dot{W}_{infra,2}$ represents the electrical power of local server fans; in liquid-cooled systems, it is the power of primary coolant pumps; in cryogenic systems, it includes the power of local compressors or expanders. The outputs are the enthalpy of the outgoing heated coolant ($\dot{H}_{2,out}$) and the unmediated heat exchange between the cooling infrastructure and the ambient environment ($\dot{Q}_{env,2}$). The First Law energy balance for this steady-flow system is:

$$\dot{Q}_{1\to2} + \dot{H}_{in,2} + \dot{W}_{infra,2} = \dot{H}_{out,2} + \dot{Q}_{env,2} \qquad (2.6)$$

Exergy Balance for Zone 2:

Correspondingly, the exergy balance must account for the flow exergy of the coolant ($\dot{E}_{flow,2}$). The macroscopic infrastructure power $\dot{W}_{infra,2}$ drawn from the grid is 100% pure exergy. The exergy balance is formulated as:

$$\dot{E}_{in,2}^{(Q)} + \dot{E}_{flow,2,in} + \dot{W}_{infra,2} = \dot{E}_{flow,2,out} + \dot{A}_{loss,2} + \dot{A}_{gen,2} \qquad (2.7)$$

Here, $\dot{A}_{loss,2}$ represents the exergy destroyed by the unmediated thermal leakage $\dot{Q}_{env,2}$ to the dead state. Crucially, $\dot{A}_{gen,2}$ represents the internal anergy generated within the cooling infrastructure itself due to fluidic viscous friction, phase-change irreversibilities, or compressor inefficiencies.

Zone 3: The Macroscopic Heat Rejection Infrastructure (Facility Level)

Zone 3 encompasses the facility-level infrastructure (e.g., Coolant Distribution Units, facility-level pumps, air handlers, cooling towers, or external cryoplant compressors) responsible for managing the heat transferred from the primary coolant loop (Zone 2). Crucially, modern high-density computing facilities do not merely reject heat to the ambient environment; they increasingly

operate as cogeneration systems, recovering high-grade waste heat for secondary applications (e.g., district heating, industrial processes, or absorption chilling).

Energy Balance for Zone 3:

This zone operates as the final thermal routing hub. The inputs are the heat transferred from the primary cooling loop ($\dot{Q}_{2\to3}$) and the macroscopic facility work ($\dot{W}_{infac,3}$) required to drive external pumps and fans, or secondary compressors. The output bifurcates into the heat recovered for useful applications ($\dot{Q}_{reuse}$) and the residual heat ultimately rejected to the ambient environment ($\dot{Q}_{3\to env}$). The steady-state energy balance is:

$$\dot{Q}_{2\to3} + \dot{W}_{infac,3} = \dot{Q}_{reuse} + \dot{Q}_{3\to env} \qquad (2.8)$$

Exergy Balance for Zone 3:

Heat rejected directly to the dead state at T_0 possesses zero residual work potential. However, the heat recovered at a useful temperature $T_{reuse} > T_0$ retains a positive exergy value, defined as $\dot{E}^{(Q)}_{reuse} = \dot{Q}_{reuse}(1 - T_0/T_{reuse})$. Thus, the exergy balance for Zone 3 is:

$$\dot{E}^{(Q)}_{2\to3} + \dot{W}_{infac,3} = \dot{E}^{(Q)}_{reuse} + \dot{A}_{gen,3} \qquad (2.9)$$

where $\dot{A}_{gen,3}$ is the anergy generated by heat exchanger inefficiencies and the final irreversible rejection of $\dot{Q}_{3\to env}$ to the dead state.

The Global Macroscopic Penalty:

To maintain strict thermodynamic consistency across the entire hierarchy, the total macroscopic infrastructure power drawn from the grid ($\dot{W}_{infra,total} = \dot{W}_{infra,2} + \dot{W}_{infac,3}$) must be evaluated against the net thermodynamic penalty. We define a **Universal Exergetic Coupling Factor, χ**, which rigorously maps the microscopic thermal penalty to the macroscopic infrastructure power. However, in systems with waste heat recovery, the recovered exergy acts as a thermodynamic offset, effectively reducing the net systemic irreversibility:

$$\dot{W}_{macro} - \dot{E}^{(Q)}_{reuse} = \chi\, \dot{A}_{total} \qquad (2.10)$$

The complete flow topology underlying Eqs. (2.3)-(2.10) is summarised in Figure 1, which tracks the continuous degradation of 100% pure electrical exergy into anergy as it cascades from the microscopic substrate to the universal waste-heat sink. Three modelling conventions adopted in the diagram should be made explicit. (i) The directional convention corresponds to the classical regime ($T_{eff} > T_0$); in cryogenic operation $\dot{Q}_{env,1}$ reverses sign and the Carnot factor $\left(1 - T_0/T_{eff}\right)$ becomes negative, as discussed above. (ii) For visual clarity, the coolant enthalpy

streams $\dot{H}_{in,2}$ and $\dot{H}_{out,2}$ of Eq. (2.6) are folded into the residual flux $\dot{Q}_{2\to3}$, and the width of the recovered-exergy arrow represents the associated heat flow $\dot{Q}_{reuse}$, since strictly $\dot{E}_{reuse}^{(Q)} = \dot{Q}_{reuse}(1 - T_0/T_{reuse}) < \dot{Q}_{reuse}$. (iii) Arrow widths illustrate the First-Law bookkeeping rather than the case-study magnitudes of Section 4.1, where $\chi = PUE - 1 \approx 0.1$ renders the infrastructure powers $\dot{W}_{infra,2}$ and $\dot{W}_{infac,3}$ substantially smaller relative to $\dot{W}_{in}$ than depicted.

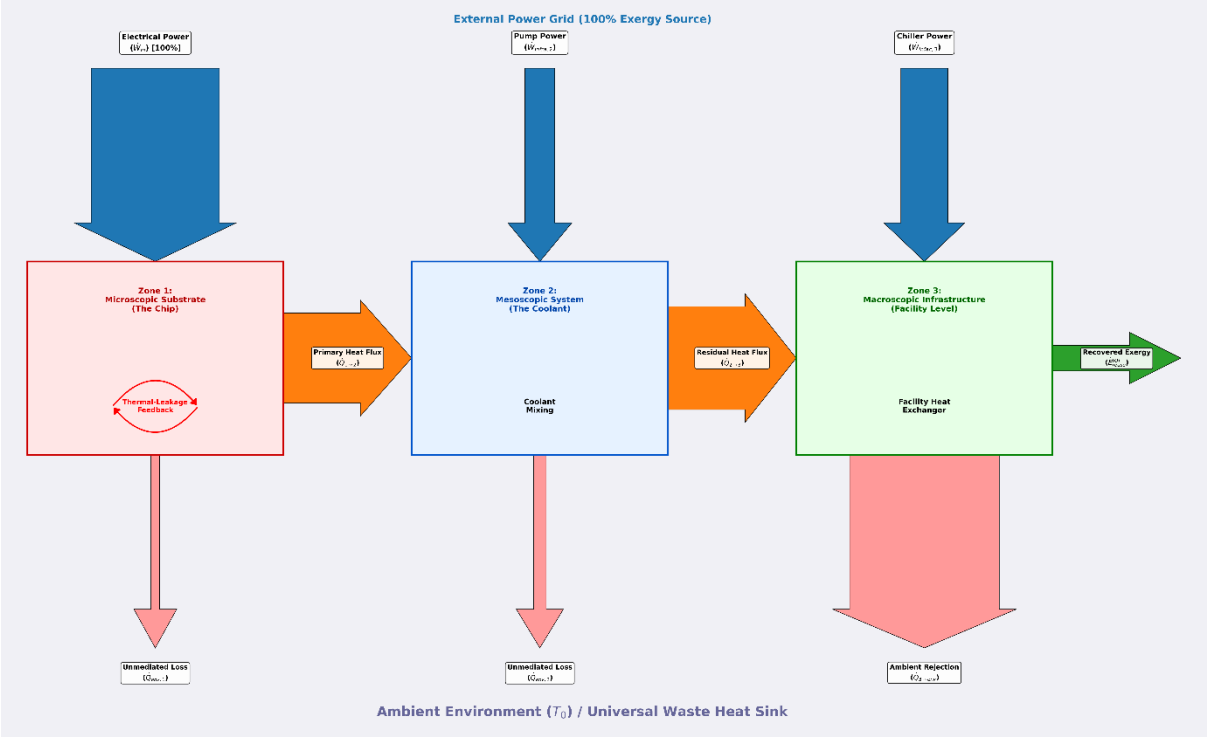

Figure 1. Schematic energy–exergy flow diagram of the three-zone thermal hierarchy.

Blue arrows: pure electrical exergy drawn from the external grid ($\dot{W}_{in}$, $\dot{W}_{infra,2}$, $\dot{W}_{infac,3}$). Orange arrows: inter-zone heat fluxes ($\dot{Q}_{1\to2}$, $\dot{Q}_{2\to3}$) carrying residual thermal exergy. Green arrow: exergy recovered at the facility heat exchanger, $\dot{E}_{reuse}^{(Q)}$. Pale-red arrows: heat thermalising directly with the ambient dead state at T_0— the unmediated losses $\dot{Q}_{env,1}$, $\dot{Q}_{env,2}$ and the final rejection $\dot{Q}_{3\to env}$. The red recirculating loop in Zone 1 denotes the non-linear thermal-leakage feedback $P_{leak}(T_{eff})$ that drives the saddle-node bifurcation of Section 3.2. Arrow widths are proportional to steady-state energy flow rates and satisfy the First-Law balances, Eqs. (2.3), (2.6) and (2.8). Magnitudes are illustrative; the directionality shown corresponds to the classical regime $T_{eff} > T_0$ (cryogenic convention and quantitative caveats are given in Section 2.1).

In classical liquid-cooled data centers, χ captures the combined fluidic and heat-exchange irreversibilities. By maximizing $\dot{E}_{reuse}^{(Q)}$ (e.g., through high-temperature two-phase cooling), the net macroscopic penalty can be drastically minimized. Conversely, in cryogenic quantum systems, heat recovery is physically impossible ($\dot{E}_{reuse}^{(Q)} = 0$), and χ captures the massive thermodynamic irreversibilities of the multi-stage refrigeration cycle, becoming astronomically large ($\chi \gg 10^3$ at

millikelvin temperatures) (see Supplementary Note S2.2; Holmes et al., 2013; Krinner et al., 2019; Cheng et al., 2024; van Staveren et al., 2025; Guha Majumdar, 2026).

2.2. Non-Equilibrium Dissipation via Onsager Formalism

The macroscopic exergy balance established in Section 2.1 treats the continuous generation rate of anergy ($\dot{A}_{total}$) as a lumped thermodynamic penalty. However, modern computation is inherently a dynamic, finite-time process. To achieve high throughput, the system must be forcefully driven away from thermal equilibrium. To rigorously quantify the continuous generation of anergy and reveal the fundamental speed-efficiency limits, we must open the macroscopic 'black box' of Zone 1 and model the non-equilibrium transport processes occurring at the nanoscale over a finite operation time τ.

In a computational system, the total rate of entropy generation is a superposition of two fundamentally distinct processes: the abstract informational entropy reduction dictated by Landauer's principle (Landauer, 1961; Parrondo et al., 2015), and the physical macroscopic dissipation driven by the coupled flow of electrical charge and heat.

To capture this dual nature, we apply the linear non-equilibrium thermodynamic framework originally developed by Onsager (Onsager, 1931) and recently extended to finite-time computational systems (Chimal-Eguia et al., 2025; Rolandi, 2026). The local rate of entropy generation, dS_{gen}/dt, is defined as the sum of the informational entropy rate and the products of physical thermodynamic fluxes (J_i) and their conjugate thermodynamic forces (X_i):

$$\frac{ds_{gen}}{dt} = \frac{1}{T_0}\frac{da_{info}}{dt} + J_e\mathrm{X}_e + J_t\mathrm{X}_t \qquad (2.11)$$

Here, the informational term is expressed via the anergy generation rate ($a_{info} = T_0 s_{info}$). Crucially, to maintain strict dimensional and thermodynamic consistency, the physical forces driving entropy production must be defined as gradients of intensive parameters. The electrical flux J_e (current density) is driven by $\mathrm{X}_e = -\nabla\phi/T$, and the thermal flux J_t (heat flux density) is driven by the gradient of inverse temperature $\mathrm{X}_t = \nabla(1/T) = -\nabla\mathrm{T}/T^2$. In a coupled physical system, these fluxes are linked via the phenomenological Onsager matrix **L**:

$$\begin{pmatrix} J_e \\ J_t \end{pmatrix} = \begin{pmatrix} L_{ee} & L_{et} \\ L_{te} & L_{tt} \end{pmatrix} \begin{pmatrix} \mathrm{X}_e \\ \mathrm{X}_t \end{pmatrix} \qquad (2.12)$$

where L_{ee} is the electrical conductivity, L_{tt} is the thermal conductivity, and the off-diagonal terms $L_{et} = L_{te}$ (by Onsager reciprocity) represent the thermoelectric cross-coupling.

To translate this local entropy production into the macroscopic loss of work potential, we apply the Gouy-Stodola theorem ($\mathrm{p}_{anergy} = T_0\, ds_{gen}/dt$). Substituting Eq. 2.12 into Eq. 2.11, the physical power dissipation density (Anergy rate) expands to:

$$\mathrm{p}_{anergy} = T_0\left(L_{ee}{X_e}^2 + 2L_{te}X_eX_t + L_{tt}{X_t}^2\right) \qquad (2.13)$$

By integrating these local entropy production rates over the chip volume V and the finite operation time τ of a computational task, we transition from the microscopic Onsager formalism to macroscopic lumped parameters. The cross-coupling terms (L_{te}, L_{et}) integrate into a macroscopic thermoelectric drag penalty, while the diagonal terms decouple the continuous Anergy generation into distinct physical power density components (Chimal-Eguia et al., 2025; Virmani & Chatterjee, 2025):

1. Dynamic Power Density (P_{dyn}): The Cost of Finite-Time Transport

To understand how the operation time τ dictates the spatial gradients X_e and X_t, we must bridge the local fluxes with macroscopic constraints. Executing a logical operation or routing data requires transporting a fixed amount of charge Q across a characteristic physical length l (e.g., a transistor channel or an interconnect wire) strictly within the time window τ.

This temporal constraint forces the system to maintain an average electrical flux (current density) proportional to the required speed: $\langle J_e \rangle \propto Q/\tau$. According to the Onsager relation, assuming the diagonal term dominates active switching, this flux is driven by the electrical force: $J_e \approx L_{ee}X_e$ (Rolandi, 2026).

Consequently, to meet the time constraint τ, the system must forcefully establish a proportional potential gradient:

$$X_{e,dyn} \approx \frac{\langle J_e \rangle}{L_{ee}} \propto \frac{1}{\tau} \qquad (2.14)$$

This reveals the physical origin of finite-time friction: the spatial gradient X_e is not a static property; it is dynamically forced by the temporal constraint τ. According to the framework of finite-time thermodynamics (Salamon & Berry, 1983), driving a system along a thermodynamic length in finite time incurs a penalty proportional to the speed. Since the local power dissipation density is the product of flux and force (J_eX_e), substituting $X_{e,dyn}$ into the Gouy-Stodola equation yields a dynamic power density that scales quadratically with the inverse of the operation time (Proesmans et al., 2020):

$$P_{dyn} = T_0L_{ee}{X_{e,dyn}}^2 = T_0L_{ee}\left(\frac{Q}{L_{ee}\tau}\right)^2 = \frac{T_0Q^2}{L_{ee}\tau^2} \qquad (2.15)$$

In modern AI accelerators, up to 80% of this dynamic energy is consumed not by logical state transitions, but by spatial data movement across the memory hierarchy and Network-on-Chip (Das et al., 2024; Tkachenko, 2025). Therefore, we explicitly decouple the classical dynamic energy parameter into the microscopic logical energy (E_{logic}) and the macroscopic spatial transport energy ($E_{transport}$), by defining the macroscopic switching energy parameter $\left(E_{logic} + E_{transport}\right)\tau_0 \equiv T_0 Q^2 / L_{ee}$, and obtain the exact finite-time scaling:

$$P_{dyn} = \frac{\left(E_{logic} + E_{transport}\right)\tau_0}{\tau^2} \qquad (2.16)$$

This quadratic inverse scaling is the hallmark of finite-time friction: computing and routing data faster incurs a massive electro-spatial penalty, driving severe transient temperature spikes. τ_0 is the intrinsic time constant. Crucially, while classical models assume τ_0 is dictated by transistor gate delays, in modern architectures it is overwhelmingly dominated by the intrinsic RC delay of the interconnects.

2. Baseline Power Density (P_{base}): The Cost of Non-Equilibrium Existence

The remaining terms in the Onsager expansion operate continuously over time, driven by static holding potentials and thermal gradients. This baseline penalty comprises three components, all of which are coupled to the thermal state of the system:

$$P_{base} = P_{leak}\left(T_{eff}\right) + P_{cond}(\nabla T) + P_{cross}(\nabla T) \qquad (2.17)$$

- Thermal Conduction (P_{cond}): The macroscopic thermal conduction loss ($P_{cond} = T_0 L_{ee} {X_{th}}^2$) through the packaging.
- Cross-Coupling Drag (P_{cross}): The off-diagonal Onsager terms explicitly capture the anergy generation rate arising from nanoscale thermoelectric cross-coupling. In modern sub-2nm 3D-ICs, extreme localized thermal gradients (∇T) actively drive parasitic charge fluxes against the electrical potential via the Seebeck and Soret effects. This creates a non-linear thermodynamic drag $P_{cross} = T_0(2L_{te}X_{el}X_{th})$ that severely compounds standard leakage as long as the chip remains hot (Li et al., 2024; Virmani & Chatterjee, 2025).
- Static Leakage (P_{leak}): Standard subthreshold leakage added as a baseline constant dependent on T_{eff} (Roy et al., 2003).

In sub-2nm architectures, the linear response regime is valid only when the relative temperature drop over the phonon mean free path λ is negligible, i.e., the thermal Knudsen condition $\lambda\,|\nabla T|/T \ll 1$ is satisfied. When extreme localized thermal gradients violate this threshold, quasi-ballistic phonon transport pushes the system toward the boundaries of the linear regime,

necessitating higher-order corrections from Extended Irreversible Thermodynamics (EIT). Because higher-order EIT corrections are strictly positive-definite, any non-linear transport phenomena will only amplify the anergy generation, meaning our derived macroscopic bifurcation limits serve as a rigorous, conservative upper bound on physical performance.

While P_{cond} and P_{crass} depend polynomially on the thermal gradient, the standard subthreshold leakage (P_{leak}) exhibits an exponential dependence on the local, which in our lumped macroscopic model is represented by the effective junction temperature temperature T_{eff} (Roy et al., 2003; Liao et al., 2017; Myeong et al., 2020; Gupta et al., 2021; Yusupov et al., 2025). We explicitly highlight this $P_{leak}(T_{eff})$ dependence because this extreme exponential sensitivity acts as the primary non-linear feedback mechanism in modern semiconductors, ultimately triggering the catastrophic thermal runaway discussed in Section 3.

The Master Equation of Finite-Time Computational Anergy:

Since continuous degradation occurs over the finite operation time τ, the total generated Anergy per operation (A_{total}) is the time-integral of these power densities, plus the fundamental discrete informational erasure limit (A_{info}). This yields the Master Equation:

$$A_{total}(\tau) = \int_0^{\tau} \left(\frac{dA_{info}}{dt} + P_{dyn} + P_{base} \right) dt = A_{info} + (P_{dyn} + P_{base})\tau \qquad (2.18)$$

This Master Equation explicitly decouples the theoretical minimum cost of logic (A_{info}) from the physical cost of moving fast across space (P_{dyn}) and the penalty of existing in a leaky, non-equilibrium state (P_{base}). It forms the rigorous foundation for evaluating the thermodynamic optimum and the saddle-node bifurcation in Section 3.

2.3. The Continuous-State Precision Penalty

To bridge the macroscopic thermal zones with microscopic logical operations, we must formalize the physical nature of information. Within the finite-time exergy framework, the irreversible erasure of a logical state fundamentally destroys the potential to perform macroscopic work. This lost work is formally defined as the generation of informational anergy (A_{info})—a discrete, fundamental fraction of the total internal anergy (A_{total}) generated in Zone 1. Thus, the thermodynamic penalty of computation is not merely 'heat' generated within the processor, but the mandatory generation of anergy paid directly to the environmental dead state at T_0.

The classical Landauer bound (Szilard, 1929; Landauer, 1961) assumes a quasistatic process in a binary system perfectly coupled to an equilibrium thermal bath. To maintain strict thermodynamic

consistency across modern hardware, we must generalize the informational entropy reduction ΔS_{info} across both discrete digital logic and continuous analog states.

Case 1: Discrete Systems (Digital Logic)

For discrete memory registers with N equiprobable microstates, the initial thermodynamic entropy of the randomized system is $S_{initial} = k_B \ln N$. Upon an irreversible logical erasure (resetting the register to a single, known reference state where $S_{final} = 0$, the informational anergy generated per operation scales logarithmically with the state space size (Parrondo et al., 2015; Manzano et al., 2024):

$$A_{info} = T_0 \Delta S_{info} = T_0 k_B \ln N \qquad (2.19)$$

This represents the fundamental Landauer limit generalized for multi-level logic (e.g., MLC/TLC flash memory or high-dimensional qudits). Erasing a state from a larger alphabet requires proportionally more thermodynamic work because a larger phase-space volume is compressed.

Case 2: Continuous Systems (Analog and Neuromorphic Computing)

Unlike discrete registers, analog and neuromorphic computing rely on continuous physical substrates (e.g., the conductance of a memristive filament). In a physical analog system, the minimum physically distinguishable precision ε_{min} is fundamentally bounded by the thermal noise floor of the local confining potential landscape (Diamantini et al., 2016; Falasco & Esposito, 2025; Berx et al., 2026).

Consider a continuous state variable x confined in a local harmonic potential well $V(x) = \kappa x^2/2$, where κ is the effective stiffness of the state confinement. At thermal equilibrium with the immediate surrounding substrate, the probability distribution of the state variable follows Boltzmann statistics governed by the strictly local hotspot temperature, T_{local} (where $T_{local} \geq T_{eff}$). According to the equipartition theorem of stochastic thermodynamics, the variance of the thermal fluctuations is:

$$\frac{1}{2}\kappa\langle x^2\rangle = \frac{1}{2}k_B T_{local} \Longrightarrow \langle x^2\rangle = \frac{k_B T_{local}}{\kappa} \qquad (2.20)$$

Thus, the absolute physical noise floor is strictly dictated by the local junction temperature:

$$\varepsilon_{min} = \sqrt{\langle x^2\rangle} = \sqrt{\frac{k_B T_{local}}{\kappa}} \qquad (2.21)$$

While the harmonic approximation provides a foundational lower bound, physical state transitions in memristive devices (e.g., oxygen vacancy migration in ReRAM) are governed by anharmonic

potential landscapes, and state stability against thermal noise is more rigorously described by Kramers' escape rate over an activation barrier (Chen & Craven, 2025). Nevertheless, the macroscopic limit remains absolute: reducing the operational precision below ε_{min} is physically impossible without actively modifying the potential landscape (increasing κ).

To bridge the continuous physical reality with the discrete thermodynamic bound, we apply a coarse-graining of the phase space. For a continuous variable with a dynamic range L, the effective number of distinguishable microstates is defined by the ratio of the total range to the precision tolerance: $N_{eff} = L/\varepsilon$.

According to the generalized Landauer principle for systems with autonomous feedback (Sagawa, 2012; Manzano et al., 2024; Wolpert et al., 2024), the total thermodynamic cost is strictly bounded by the mutual information required to maintain the state. Taking the formal continuous limit, the cumulative informational anergy for analog workloads becomes:

$$A_{info} = T_0 \Delta S_{info} = T_0 k_B \ln\left(\frac{L}{\varepsilon}\right) \qquad (2.22)$$

The Precision-Temperature Trap

Equation (2.22) mathematically defines the Continuous-State Precision Penalty. This derivation reveals a fundamental dual-temperature dependence: the physical thermal noise ε_{min} (Eq. 2.21) is strictly governed by the local hot substrate at T_{eff}, yet the irreversible loss of work potential (Anergy, Eq. 2.22) resulting from compressing this noise is anchored to the ambient dead-state environment at T_0 (via the Gouy-Stodola theorem).

This dual-temperature dependence exposes the physical origin of the 'Precision-Temperature Trap'. To achieve higher precision (smaller $\varepsilon \to \varepsilon_{min}$), the system must artificially suppress the thermal noise floor. From Equation (20), this requires increasing the confining stiffness quadratically ($\kappa \propto 1/\varepsilon^2 \propto N^2$).

Crucially, in physical devices like ReRAM, this state stiffness κ is directly proportional to the macroscopic logical switching energy, E_{logic} (which manifests as the Joule heating applied during SET/RESET programming pulses), required to form a thicker, more rigid, and noise-resistant conductive filament (Ielmini & Pedretti, 2025). Since the minimum distinguishable precision scales inversely with the number of states ($\varepsilon \propto 1/N$), and the thermal noise floor is given by $\varepsilon_{min} = \sqrt{T_{local} k_B/\kappa}$, maintaining state distinguishability requires the confining stiffness κ to scale quadratically with N. Consequently, pushing analog systems toward digital-like precision (high N) requires an exponential explosion in this logical energy:

$$E_{logic} \propto \kappa \propto N^2 T_{eff} \qquad (2.23)$$

As illustrated in Figure 2, while discrete digital systems follow the logarithmic Landauer scaling, analog substrates exhibit a severe non-linear Precision Penalty. As precision demands increase, the structural anergy required to maintain tight potential wells against thermal fluctuations at T_{local} drives the analog thermodynamic cost into a physically unreachable regime, establishing strict fundamental boundaries on the analog computing advantage.

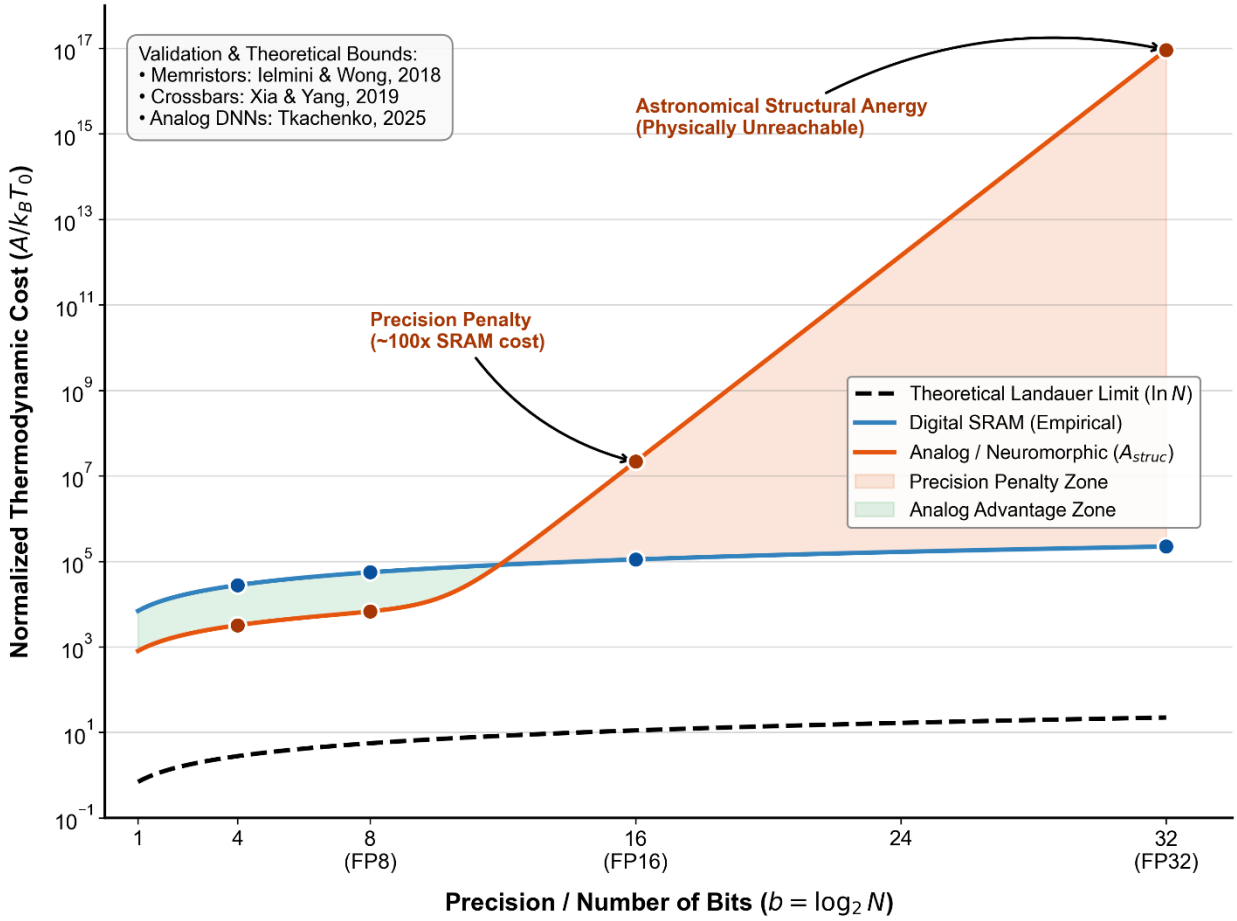

Figure 2. The generalized Landauer limit across computing paradigms.

The semi-logarithmic plot demonstrates the scaling of thermodynamic cost (informational entropy) as a function of the effective number of distinguishable states N, highlighting the structural 'Precision Penalty' inherent to analog and neuromorphic substrates.

3. The Speed-Efficiency Paradox

While Section 2 establishes the fundamental components of computational anergy—informational, dynamic, and static—these components exhibit inherently opposing dependencies on the operation time τ. Dynamic transport heavily penalizes fast operations ($\propto 1/\tau^2$), whereas static leakage penalizes slow operations ($\propto \tau$). This temporal conflict sets the stage for a fundamental optimization problem.

The classical interpretation of Landauer's principle suggests that computation becomes perfectly efficient as it becomes infinitely slow ($\tau \to \infty$). However, our generalized finite-time exergy framework reveals a fundamental thermodynamic paradox (Curzon & Ahlborn, 1975; Lipka-Bartosik & Perarnau-Llobet, 2025). If a computation is performed infinitely slowly, the static maintenance cost $P_{maint}\tau$ diverges to infinity. This is particularly catastrophic in deep-submicron CMOS due to quantum tunneling (Roy et al., 2003; Chaudhry & Kumar, 2004). Consequently, physical computing systems cannot operate at the quasistatic Landauer limit; they are forced to find an optimal finite operation time τ_{opt} that balances dynamic friction against static leakage.

3.1. Implicit Thermal-Leakage Feedback

To determine the true optimal operation time τ_{opt}^{*} that minimizes the local electrical work W_{elec} dissipated within the microscopic control volume, we cannot treat the effective temperature T_{eff} as a static parameter. In modern high-density architectures, T_{eff} is dynamically coupled to both the total dissipated electrical work W_{elec} and the operation time τ (which dictates the transient power density).

According to the Gouy-Stodola theorem, the total electrical work W_{elec} (which enters the system as pure exergy) is ultimately dissipated as heat at the effective junction temperature T_{eff}. The corresponding entropy generation is $\Delta S = W_{elec}\,/T_{eff}$. Therefore, the total anergy generated is $A_{total} = T_0 \Delta S = T_0\left(W_{elec}\,/T_{eff}\right)$. Rearranging this fundamental balance yields the implicit thermodynamic equilibrium constraint that must be strictly satisfied at any given operation speed. We must treat the energy balance as an implicit constraint. We define the implicit function $F(W_{elec}, \tau) = 0$, which represents the fundamental thermodynamic equilibrium that must be strictly satisfied at any given operation speed:

$$F(W_{elec}, \tau) \equiv \frac{T_{eff}(W_{elec}, \tau)}{T_0} A_{total}\left(\tau, T_{eff}(W_{elec}, \tau)\right) - W_{elec} = 0 \qquad (3.1)$$

By applying the Implicit Function Theorem, the true rate of change of the required electrical work is given by the ratio of its partial derivatives: $dW_{elec}/d\tau = -(\partial F/\partial \tau)(\partial F/\partial W_{elec})$. To unpack this derivative, we apply the multivariate chain rule, isolating four fundamental thermodynamic response factors:

1. $\alpha \equiv \left(\partial T_{eff}/\partial \tau\right)_{W_{elec}} < 0$: **Transient Thermal Response Rate.** Compressing the workload into a shorter time τ spikes the instantaneous dynamic power density, leading to a transient temperature surge.

2. $\zeta \equiv \left(\partial T_{eff}/\partial W_{elec}\right)_{\tau} > 0$: **Steady-State Thermal Resistance.** This coincides with the effective thermal resistance R_{th_eff} of the chip-package interface.
3. $\gamma \equiv (\partial A_{total}/\partial \tau)_{T_{eff}} = -P_{dyn} + P_{base}$: **Direct Temporal Anergy Trade-off.** The classical balance between dynamic friction and static leakage at a constant temperature.
4. $\delta \equiv \left(\partial A_{total}/\partial T_{eff}\right)_{\tau} \gg 0$: **Area-normalized Thermal-Leakage Feedback Factor.** Driven primarily by the exponential dependence of subthreshold leakage currents on temperature.

Evaluating the Implicit Function Theorem using these factors yields the exact rate of change of the required electrical work:

$$\frac{dW_{elec}}{d\tau} = \frac{T_{eff}\gamma + \left(A_{total} + T_{eff}\delta\right)\alpha}{T_0 - \zeta\left(A_{total} + T_{eff}\delta\right)} \qquad (3.2)$$

Assuming the cooling infrastructure maintains thermal stability (denominator > 0), the thermodynamic optimum occurs when the numerator equals zero. Dividing by T_{eff} and substituting $\alpha = -|\alpha|$ to make the physical directionality transparent, we obtain the exact optimality condition:

$$-P_{dyn} + P_{base} - P_{penalty} = 0 \qquad (3.3)$$

Where

$$P_{penalty} \equiv |\alpha|\left(\tau\frac{\partial P_{base}}{\partial T_{eff}} + \frac{A_{total}}{T_{eff}}\right) = |\alpha|\left(\delta + \frac{A_{total}}{T_{eff}}\right) \qquad (3.4)$$

is the Thermal Brake Penalty. This term encapsulates the non-linear thermodynamic drag: computing faster triggers a transient temperature spike (driven by $|\alpha|$), which exponentially amplifies the static leakage (driven by δ).

To formalize this thermodynamic balance, we define the **Exact Effective Dissipation**, D_{exact}, as the net thermodynamic friction remaining after the Thermal Brake has partially canceled the baseline leakage:

$$D_{exact} \equiv P_{base} - P_{penalty} \qquad (3.5)$$

By substituting the explicit finite-time scaling of dynamic power (Eq. 2.16) (Proesmans et al., 2020), into the exact optimality condition (Eq. 3.5), we can algebraically isolate the true optimal operation time τ^*_{opt}:

$$\tau^*_{opt} \approx \sqrt{\frac{\left(E_{logic} + E_{transport}\right)\tau_0}{D_{exact}}} \qquad (3.6)$$

We use an approximate equality (≈) to rigorously acknowledge that Equation (3.6) is an implicit, self-consistent relationship. The variable τ remains present on the right-hand side within the $P_{penalty}$ term (which contains the transient response rate α). However, because the dynamic friction term scales as $1/\tau^2$ and overwhelmingly dominates the temporal forces in the high-frequency limit, isolating it yields a highly accurate structural formula that captures the leading-order behavior of the system (Salamon & Berry, 1983).

Equation 3.6 is the central analytical result of the microscopic optimization. It proves that the optimal speed of computation is dictated by two opposing forces defining D_{exact}:

1. **The 'Race to Sleep' (Denominator Inflation):** As transistors scale down, static leakage P_{base} grows exponentially. inflating D_{exact} and forcing τ^*_{opt} to shrink (driving the system to compute faster).
2. **The 'Thermal Brake' (Denominator Reduction):** Computing faster causes a transient temperature spike, triggering the thermal-leakage feedback. Because $P_{penalty}$ is subtracted from the denominator, it shrinks D_{exact}, forcing τ^*_{opt} to increase.

Crucially, if the thermal braking term ever equals the static dissipation ($P_{penalty} \to P_{base}$), the effective dissipation denominator D_{exact} vanishes. This mathematical singularity indicates that the system can no longer find a steady-state thermal equilibrium, providing a rigorous thermodynamic explanation for the empirical 'thermal throttling' universally observed in modern 3D-ICs (Virmani & Chatterjee, 2025).

3.2. The Master Equation and Saddle-Node Bifurcation

Equation (3.2) reveals that the stability of the thermal equilibrium is strictly governed by its denominator. We explicitly define this denominator as the **Thermodynamic Stability Margin**, $\mathcal{M} \equiv T_0 - \zeta\left(A_{total} + T_{eff}\delta\right)$.

To rigorously prove the system's thermal trajectory, we evaluate the total transient temperature response

$$\frac{dT_{eff}}{d\tau} = \alpha + \zeta\frac{dW_{elec}}{d\tau} \qquad (3.7)$$

Substituting Eq. (3.2) into Eq. (3.7) and expressing it over the common denominator $\mathcal{M}$ reveals an exact algebraic cancellation of the non-linear thermal-leakage feedback terms (see

Supplementary Note S2.2 for the full step-by-step expansion). The trajectory simplifies to a strictly defined, closed-form equation:

$$\frac{dT_{eff}}{d\tau} = \frac{\alpha T_0 + \zeta T_{eff}\gamma}{\mathcal{M}} \qquad (3.8)$$

Equation (3.8) proves that the transient trajectory of the effective temperature is strictly decoupled from the non-linear leakage amplification in the numerator. The system's stability is entirely governed by the Thermodynamic Stability Margin, $\mathcal{M}$. From an algebraic perspective, the condition $\mathcal{M} \to 0$ indicates a singularity in the implicit function (Eq. 3.1), representing a boundary collapse of the steady-state solution space. In the context of dynamical systems, this singularity manifests as a catastrophic saddle-node bifurcation (Strogatz, 2015; Falasco & Esposito, 2025). At this critical threshold, the stable and unstable thermal equilibrium points collide and annihilate; any infinitesimal increase in dissipated work leads to a divergent temperature increase, formally defining the absolute physical speed limit of the computing substrate, a critical reliability bottleneck recently identified in dense 3D-ICs (Kim et al., 2026).

To extract a closed-form analytical expression for this absolute physical speed limit, we substitute the explicit physical definitions of $\zeta = R_{th\_eff}/\tau$, $A_{total} = A_{info} + E_{dyn}(\tau_0/\tau) + P_{base}\tau$, and $\delta = \tau(\partial P_{base}/\partial T_{eff})$ into the bifurcation condition $\mathcal{M} = 0$. Solving the resulting quadratic equation for τ yields the absolute thermodynamic speed limit of computation, $\tau_{runaway}$:

$$\tau_{runaway} = \frac{A_{info} + \sqrt{{A_{info}}^2 + 4C_{res}(E_{logic} + E_{transport})\tau_0}}{2C_{res}} \qquad (3.9)$$

where $C_{res} \equiv T_0/R_{th_{eff}} - P_{base} - T_{eff}(\partial P_{base}/\partial T_{eff})$ is the residual thermal conductance capacity of the package.

Equation (3.9) establishes the Landauer Speed Penalty. It proves that the absolute frequency limit ($f_{runaway} = 1/\tau_{runaway}$) is fundamentally coupled to the microscopic informational cost A_{info}. The discrete informational anergy generated per logical erasure adds directly to the baseline thermal load, effectively 'choking' the dynamic power growth and forcing the system to undergo bifurcation at a lower frequency. Consequently, the absolute physical limit of computation is no longer dictated by transistor gate delays, but by the topological proximity to this saddle-node bifurcation boundary in the macroscopic phase space (Figure 3) (Falasco & Esposito, 2025).

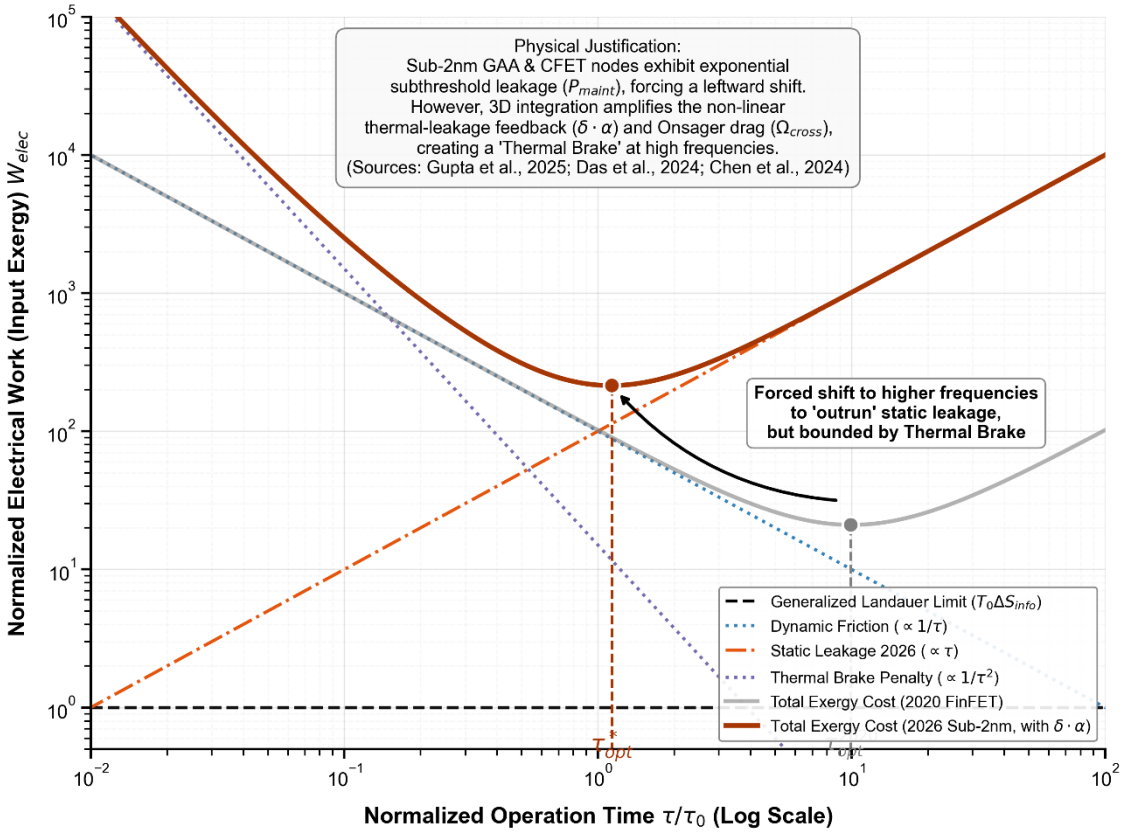

Figure 3. The Speed-Efficiency Paradox.

Normalized electrical work is plotted against normalized operation time on a log-log scale. The U-shaped curves highlight the leftward shift of the optimal operation point in sub-2nm architectures, driven by the necessity to 'outrun' exponentially increasing static leakage.

4. Detailed Case Studies

To demonstrate the predictive power of the Finite-Time Computational Exergy framework, we apply the Master Equation and the Unified Optimality Conditions to the two dominant paradigms of modern AI hardware: digital sub-2nm 3D-ICs and analog memristive crossbars. By explicitly calculating the partitioning of Anergy and the shift in optimal operation times, we quantitatively isolate the distinct physical bottlenecks dictating the absolute speed limits of each architecture.

4.1. The Digital Limit: Sub-2nm 3D-ICs and the Thermal Brake

Modern digital AI accelerators utilizing sub-2nm Complementary FET (CFET) nodes and 3D heterogeneous integration operate strictly in the dissipative regime (Chen, S. et al., 2024). The adoption of Backside Power Delivery Networks (BSPDN) necessitates extreme wafer thinning, which catastrophically degrades lateral heat spreading and inflates the transient thermal impedance, $R_{th_{eff}}$ (Xie & Wei, 2024; Virmani & Chatterjee, 2025; Kim & Kwon, 2026).

To anchor our Master Equation in empirical hardware constraints, we evaluate the exact optimality condition for a 1 mm² active tensor core block using projected 2026 metrics. Due to severe interconnect resistance scaling, spatial data transport dominates the dynamic energy budget. Calculating the transport energy via the physical charging of the parasitic capacitance $E_{transport} = 1/2\, C_{int,total} {V_{DD}}^2$ (where $C_{int,total}$ is the total switched parasitic capacitance density of the interconnects and V_{DD} is the supply voltage), yields $8 \times 10^{-8}\ J/mm^2$. Combined with the logic switching energy, this produces a macroscopic switching parameter of $(E_{logic} + E_{transport})\tau_0 = 10^{-18}\, J \cdot s/mm^2$ (Gupta et al., 2021). The total baseline static dissipation—comprising subthreshold leakage, thermal conduction, and Seebeck-driven cross-coupling drag—is $P_{base} = 8.5\ W/mm^2$ (Virmani & Chatterjee, 2025). Evaluating the rigorous continuous volumetric integral over the 3D power-temperature profile yields a true entropy-flux-preserving effective junction temperature of $T_{eff} \approx 345.5\ K$ (Zhu et al., 2025). Finally, we adopt a median effective thermal resistance of $R_{th_{eff}} = 0.8\ K \cdot mm^2/W$, characteristic of BSPDN architectures lacking bulk silicon heat spreaders (Xie & Wei, 2024).

The Naive Isothermal Optimum

If we temporarily ignore the transient thermal feedback ($P_{penalty} = 0$), classical finite-time thermodynamics dictates that the system simply balances dynamic friction against the baseline leakage at a constant temperature. Under this naive isothermal assumption, the optimal operation time is:

$$t_{iso} = \sqrt{\frac{(E_{logic} + E_{transport})\tau_0}{P_{base}}} \approx 0.343\ ns$$

This predicts a highly optimistic optimal clock frequency of $f_{iso} \approx 2.92\ GHz$.

The Thermal Brake and Dynamical Convergence

In physical reality, the exact optimality condition derived in Section 3.1 is a transcendental equation. The system's true operational speed is governed by a non-linear dynamical feedback loop, where the operation time τ dictates the transient temperature spike, which in turn exponentially amplifies the leakage that dictates the optimal τ.

If the processor attempts to operate at the naive isothermal optimum of 2.92 GHz, the massive transient heating triggers a severe Thermal Brake penalty of $P_{penalty}(t) \approx 4.6\ W/mm^2$. This consumes over half of the package's effective dissipation capacity, forcing the system to violently throttle down to 1.98 GHz ($\tau \approx 0.506\ ns$). However, at this lower frequency, the transient thermal

impedance relaxes, the temperature drops, and the system 'sees' available thermal headroom, allowing it to speed up again.

By iteratively solving the implicit Master Equation (see Supplementary Note S4 for the full 10-step numerical convergence), we find that this oscillatory 'throttling' perfectly dampens, converging to a stable thermodynamic attractor at $\boldsymbol{\tau^*_{opt} \approx 0.424\ ns}$. This yields a true sustained clock frequency of $\boldsymbol{f^*_{opt} \approx 2.36\ GHz}$.

At this exact equilibrium, the Thermal Brake continuously consumes 34.5% of the package's effective dissipation capacity ($P_{penalty} \approx 2.94\ W/mm^2$, leaving $D_{exact} \approx 5.56\ W/mm^2$). This rigorously proves that the empirical 'thermal throttling' universally enforced by modern 3D-IC power management units is not merely a software-defined performance penalty; it is a fundamental dynamical systems adaptation required to maintain thermodynamic stability.

The Absolute Bifurcation Limit and Landauer Speed Penalty

To quantify the absolute physical speed limit where the system undergoes a saddle-node bifurcation (thermal runaway), we evaluate the Thermodynamic Stability Margin ($\mathcal{M} \to 0$). The residual thermal capacity of the packaging, after accounting for linearized leakage amplification, is $C_{res} \approx 257.9\ W/mm^2$.

Crucially, solving the exact bifurcation condition yields a quadratic equation (Eq. 3.7), proving that the absolute speed limit is fundamentally coupled to the microscopic informational cost A_{info}. Assuming a dense logic block erasing $\sim 10^7$ bits per clock cycle per mm^2, the informational anergy is $A_{info} \approx 2.8 \times 10^{-14}\ J/mm^2$.

$$\tau_{runaway} \approx 6.226 \times 10^{-11}\ s = 0.06226\ ns.$$

This establishes an absolute, uncrossable physical wall at $f_{runaway} \approx 16.1\ GHz$. The presence of A_{info} in the numerator mathematically enforces the Landauer Speed Penalty: the fundamental thermodynamic cost of information erasure directly adds to the baseline thermal load, effectively 'choking' the dynamic power growth and forcing the system to undergo bifurcation at a lower frequency than classical transport models would predict. Attempting to drive the clock beyond 16.06 GHz results in a divergent temperature increase, as no steady-state thermal equilibrium can exist.

The Macroscopic Exergetic Cost: The Optimum vs. The Runaway Limit

To fully grasp the physical implications of the Speed-Efficiency Paradox, we must evaluate these frequency limits not merely in terms of local heat generation, but through the lens of the total

macroscopic exergy destroyed per logical operation (Ex_{opt}). According to the generalized exergy balance (Eq. 2.10), the total work drawn from the grid is the product of the local anergy generated and the macroscopic penalty multiplier $\Lambda(T_{eff}, \chi)$. Assuming a state-of-the-art facility cooling infrastructure with a Power Usage Effectiveness (PUE) of 1.1 (characteristic of highly optimized hyperscale data centers), the exergetic inefficiency factor is $\chi = PUE - 1 \approx 0.1$ (Shah et al., 2008; Ebrahimi et al., 2014; Dincer & Rosen, 2020). Using this, we can quantitatively compare the thermodynamic cost at the optimal and runaway frequencies.

At the **exergetically optimal frequency ($f^*_{opt} \approx 2.36\ GHz$)**, the system operates at a stable thermodynamic attractor. The dynamic power is $P_{dyn} \approx 5.56\ W/mm^2$, and the total power density (including baseline leakage) is $P_{total} \approx 14.1\ W/mm^2$. The local energy dissipated per operation is $E_{op} = P_{total}\tau^*_{opt} \approx 5.96\ nJ/mm^2$. At the stable effective temperature of $T_{eff} \approx 345.5\ K$., the macroscopic penalty multiplier is relatively low ($\Lambda \approx 1.20$). Consequently, the total macroscopic exergy destroyed per operation is:

$$Ex_{op}(f^*_{opt}) = E_{opt}\Lambda \approx 7.13\ nJ/mm^2$$

This represents the absolute minimum thermodynamic cost required to sustain computation in this specific 3D-IC architecture.

Conversely, attempting to push the processor to the **absolute bifurcation limit ($f_{runaway} \approx 16.1\ GHz$)** results in catastrophic exergetic losses. At this frequency, the dynamic power explodes to $P_{dyn} \approx 257.6\ W/mm^2$. The localized temperature spikes violently, pushing the total power density to the absolute cooling limit of the package ($P_{total} \approx 366\ W/mm^2$). The local energy dissipated per operation skyrockets to $E_{op} \approx 22.8\ nJ/mm^2$. Furthermore, because the effective junction temperature approaches >500 K at the bifurcation boundary, the cooling infrastructure is severely strained, inflating the macroscopic penalty multiplier to $\Lambda \approx 2.10$. The total macroscopic exergy destroyed per operation becomes:

$$Ex_{runaway}(f_{runaway}) = E_{opt}\Lambda \approx 47.88\ nJ/mm^2$$

Conclusion:

Operating at the thermal runaway limit destroys nearly **6.7 times more exergy per operation** than operating at the thermodynamic optimum. This quantitative divergence rigorously proves that pushing for maximum clock speed in sub-2nm 3D-ICs is thermodynamically ruinous (Cobos-Murcia, 2026). The true goal of post-Moore hardware scaling must not be the blind pursuit of higher frequencies, but the minimization of the macroscopic exergy cost. To achieve higher throughput without incurring this massive exergetic penalty, engineers must focus entirely on

reducing the effective thermal resistance $R_{th_{eff}}$ (e.g., via microfluidics), which physically shifts the exergetic optimum f_{opt}^{*} to a higher frequency while maintaining thermodynamic stability (Virmani & Chatterjee, 2025).

4.2. The Analog Limit: ReRAM Crossbars and the Precision-Temperature Trap

Analog in-memory computing (CiM) utilizing memristive crossbars (e.g., ReRAM) promises to bypass the von Neumann bottleneck by performing matrix-vector multiplications directly within the memory array (Ielmini & Wong, 2018; Ielmini & Pedretti, 2025). However, applying our Continuous-State Precision Penalty (Section 2.3) reveals a severe thermodynamic barrier preventing analog substrates from achieving digital-like precision.

As established in Section 2.3 (Eq. 2.23), achieving an N-state multi-level cell (MLC) in a physical ReRAM device requires the macroscopic logical switching energy (E_{logic}) to scale quadratically with the number of states ($E_{logic} \propto N^2 T_{eff}$) to overcome the thermal noise floor.

Applying this Precision-Temperature Trap to a projected 2026 ReRAM array reveals a severe bottleneck. Attempting to achieve 8-bit precision ($N = 256$) requires a filament stiffness $256^2 = 65{,}536$ times greater than a binary state. Delivering this massive E_{logic} via SET/RESET pulses induces severe localized Joule heating (Swoboda et al., 2023), driving the local hotspot temperature T_{local} far above the bulk T_{eff} (Swoboda et al., 2023). Recent demonstrations of ultra-low-voltage resistive switching in ZrOx/TaOx bilayer structures suggest that material-level engineering can partially mitigate this energy overhead (Wiśniewski et al., 2025; van Gerven, 2026; Lee et al., 2026). Because the thermal noise floor itself expands with temperature ($\varepsilon_{min} \propto \sqrt{T_{local}}$), the system must inject even more programming energy to maintain the potential well, creating a runaway positive feedback loop (Ielmini & Pedretti, 2025).

Furthermore, in dense crossbar arrays, this localized heating is compounded by spatial thermal crosstalk and parasitic line resistances (sneak-path currents), which act as continuous sources of non-equilibrium dissipation (P_{cross}) (Ielmini & Wong, 2018; Xia & Yang, 2019; Kim & Yang, 2024; Pande et al., 2024).

This thermal divergence dictates a hard physical speed limit. While the absolute thermodynamic bifurcation ($f_{runaway}$) occurs when the required programming energy diverges to infinity, physical crossbars are strictly bounded by a material degradation limit ($T_{max} \approx 400\ K$). As detailed in Supplementary Note S3, evaluating this Material Limit reveals that the maximum practical clock frequency (f_{max}) collapses from $\sim 10.8\ GHz$ at 4-bit precision to a mere ~75 MHz at 8-bit precision. Consequently, while analog ReRAM exhibits exceptional thermodynamic efficiency at

low precisions (1-4 bits), pushing it toward 8-bit precision causes the structural anergy to diverge, making it thermodynamically and practically inferior to digital SRAM at high throughputs.

The analog Precision-Temperature Trap demonstrates that the thermodynamic cost of continuous-state computing is highly asymmetric. While analog crossbars offer unparalleled exergy efficiency for low-precision tasks (e.g., 4-bit weights in neural networks), forcing them to emulate digital-like precision (8-bit and beyond) shifts the physical bottleneck from spatial transport to localized Joule heating. This exponential structural penalty rigorously proves that the future of analog in-memory computing lies in heterogeneous, mixed-precision architectures rather than as a universal replacement for digital logic (Mehonic et al., 2024; Shooshtari et al., 2025).

4.3. Discussion and Methodological Limitations

The application of the Finite-Time Computational Exergy framework to sub-2nm architectures reveals a severe thermodynamic bottleneck. However, while the proposed framework provides a rigorous macroscopic bound, several physical simplifications must be acknowledged for future refinement:

1. Non-Linear Transport at Extreme Gradients: The derivation relies on the linear Onsager formalism (Eq. 2.12). In sub-2nm nodes, extreme localized thermal gradients may push electron-phonon interactions beyond the linear response regime, necessitating higher-order corrections from Extended Irreversible Thermodynamics (EIT) (Jou et al., 2010).
2. Spatial Thermal Crosstalk: The use of a lumped effective thermal resistance ($R_{th_{eff}}$) simplifies the highly heterogeneous 3D heat flow. In dense 3D-ICs, lateral thermal crosstalk between adjacent cores creates dynamic, spatially dependent thermal bottlenecks not fully captured by a single 1D impedance parameter (Moore & Shi, 2014; Guo et al., 2025).
3. Macroscopic Chip-Level Scaling: The empirical parameterization is based on a 1 mm² tensor core block. Scaling this to a full reticle-size accelerator (e.g., >800 mm²) is non-trivial. Global power delivery constraints and the necessity to power-gate inactive regions ('dark silicon') will introduce additional macroscopic boundary conditions that further constrain the global thermodynamic optimum (Esmaeilzadeh et al., 2011).

5. Hardware Engineering Imperatives

The Finite-Time Exergy framework mathematically proves that the absolute limits of modern computation are no longer dictated by transistor scaling, but by macroscopic thermodynamic bottlenecks. To sustain performance scaling without triggering the saddle-node bifurcation

(thermal runaway) or falling into the precision-temperature trap, hardware engineering must pivot toward the following paradigm-specific imperatives:

Imperative for Digital 3D-ICs: Minimizing ζ via Active Microfluidics

Equation (3.9) demonstrates that the runaway frequency $f_{runaway}$ is strictly bounded by the residual thermal capacity C_{res}, which is inversely proportional to the effective thermal resistance $R_{th_{eff}}$. The current industry trajectory of relying on Backside Power Delivery Networks (BSPDN) exacerbates this by requiring extreme wafer thinning, effectively removing the silicon heat spreader (Xie & Wei, 2024). To shift the thermodynamic optimum f_{opt}^{*} to higher frequencies, engineers must abandon passive conduction and integrate active 3D microfluidic cooling directly into the interposer or between logic tiers (Zhang et al., 2023; Heydari et al., 2024). By drastically reducing $R_{th_{eff}}$, microfluidics suppresses the transient thermal response rate α, delaying the onset of the Thermal Brake (Rathore et al., 2023; Zhang et al., 2023).

Imperative for Analog CiM: Heterogeneous Integration and Photonic Substrates

Because the analog Precision-Temperature Trap is a fundamental consequence of the fluctuation-dissipation theorem, attempting to engineer high-precision (8-bit+) memristive crossbars faces prohibitive thermodynamic barriers. Instead, architectures must embrace heterogeneous integration: utilizing analog crossbars exclusively for massive, low-precision (2-4 bit) multiply-accumulate operations, while offloading high-precision accumulation and activation functions to digital cores (Mehonic et al., 2024).

Alternatively, to fundamentally bypass the $k_B T$ noise floor, analog computation must transition to photonic crossbars. In optical neural networks, the physical state is encoded in the phase or amplitude of light (Aghaee Rad et al., 2025). The fundamental noise floor is dictated by quantum shot noise rather than thermal fluctuations, as optical frequencies ($hv \gg k_B T$) decouple the informational state from the ambient thermal bath (Shastri et al., 2020; Aghaee Rad et al., 2025). This allows photonic accelerators to scale to higher precisions without incurring the exponential structural anergy penalty inherent to electronic memristors.

6. Conclusion

By unifying Landauer's principle with finite-time thermodynamics and macroscopic exergy analysis, we have resolved the Speed-Efficiency Paradox of modern computing. We demonstrated that the continuous generation of anergy—driven by dynamic transport and static leakage—creates a non-linear thermal-leakage feedback loop that strictly bounds the maximum operational speed of any dissipative computing substrate.

Our exact analytical solution for the saddle-node bifurcation proves that empirical phenomena like 'thermal throttling' are not mere engineering artifacts, but fundamental thermodynamic imperatives required to maintain system stability. Furthermore, by generalizing the informational entropy reduction to continuous states, we exposed the Precision-Temperature Trap, establishing strict physical boundaries on the viability of analog computing. Ultimately, the future of post-Moore hardware scaling relies not on the blind pursuit of higher clock frequencies or denser integration, but on the rigorous minimization of macroscopic exergy destruction through advanced thermal packaging and hybrid electro-photonic architectures. In essence, the absolute physical limits of computation are no longer dictated by the quantum mechanics of the transistor, but by the macroscopic thermodynamics of the cooling infrastructure.

Acknowledgments and disclosures

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Competing Interests

The author declares that there are no competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

The empirical analysis in this study is based on data and parameters derived from publicly available literature cited in the references. No new datasets were generated or analysed during the current study.

Supplementary Materials

Supplementary Note S2: Towards a generalized exergy balance for finite-time calculations

S2.1. Entropy-Flux-Preserving Effective Junction Temperature (T_{eff})

In modern 3D-ICs, the local power density $p(\boldsymbol{r})$ $(in\ W/m^3)$ and the local temperature $T(\boldsymbol{r})$ $(in\ Kelvin)$ vary drastically across the spatial volume V of the chip. According to classical thermodynamics, the local rate of entropy generation due to heat transfer from a differential volume element to the ambient is proportional to the local entropy flux (Callen, 1985). Under

steady-state conditions, assuming no internal entropy sinks, the integral of the local entropy flux over the volume yields the total internal entropy generation rate (dS_{gen}) of the chip:

$$dS_{gen} = \int_V \frac{p(\boldsymbol{r})}{T(\boldsymbol{r})} dV \qquad (S2.1)$$

In macroscopic exergy analysis, we treat the chip as a single lumped node dissipating the total power $P_{total} = \int_V p(\boldsymbol{r})dV$ at a single effective temperature T_{eff}. To ensure that this lumped model strictly preserves the Second Law of Thermodynamics, the macroscopic internal entropy generation must exactly equal the integrated local entropy generation (Seifert, 2025):

$$\frac{P_{total}}{T_{eff}} = dS_{gen} \Longrightarrow T_{eff} = \frac{\int_V p(\boldsymbol{r})dV}{\int_V \frac{p(\boldsymbol{r})}{T(\boldsymbol{r})} dV} \qquad (S2.2)$$

This rigorous definition ensures that T_{eff} is mathematically anchored to the thermal mass where the vast majority of the entropy is actually generated, providing a thermodynamically consistent bridge between microscopic non-equilibrium hotspots and macroscopic heat rejection.

Supplementary Note S3: Towards the Speed-Efficiency Paradox

S3.1. Algebraic Cancellation of Non-Linear Feedback in the Transient Trajectory

To derive Equation (3.8), we start with the total transient temperature response (Eq. 3.7):

$$\frac{dT_{eff}}{d\tau} = \alpha + \zeta \frac{dW_{elec}}{d\tau} \qquad (S3.1)$$

Substituting the exact rate of change of electrical work (Eq. 3.2) into this equation yields:

$$\frac{dT_{eff}}{d\tau} = \alpha + \zeta \frac{T_{eff}\gamma + \left(A_{total} + T_{eff}\delta\right)\alpha}{T_0 - \zeta\left(A_{total} + T_{eff}\delta\right)} \qquad (S3.2)$$

Let the Thermodynamic Stability Margin be the common denominator: $\mathcal{M} \equiv T_0 - \zeta\left(A_{total} + T_{eff}\delta\right)$. Expressing the entire equation over $\mathcal{M}$:

$$\frac{dT_{eff}}{d\tau} = \frac{\alpha\left[T_0 - \zeta\left(A_{total} + T_{eff}\delta\right)\right] + \zeta T_{eff}\gamma + \zeta\left(A_{total} + T_{eff}\delta\right)\alpha}{\mathcal{M}} \qquad (S3.3)$$

Expanding the terms in the numerator:

$$Numerator = \alpha T_0 - \zeta\left(A_{total} + T_{eff}\delta\right)\alpha + \zeta T_{eff}\gamma + \zeta\left(A_{total} + T_{eff}\delta\right)\alpha \qquad (S3.4)$$

The non-linear leakage feedback terms ($-\zeta(A_{total} + T_{eff}\delta)\alpha$ and $\zeta(A_{total} + T_{eff}\delta)\alpha$) perfectly cancel each other out. The trajectory simplifies exactly to:

$$\frac{dT_{eff}}{d\tau} == \frac{\alpha T_0 + \zeta T_{eff}\gamma}{\mathcal{M}} \qquad (S3.5)$$

This rigorous cancellation proves that the transient thermal trajectory is decoupled from the leakage amplification δ in the numerator, making the system's stability strictly dependent on the singularity condition $\mathcal{M} \to 0$.

Supplementary Note S4: Detailed Parameterization and Numerical Convergence for 2026 Architectures

S4.1. Empirical Parameterization and Bifurcation Analysis for Sub-2nm CFETs

To rigorously evaluate the exact optimality condition (Eq. 3.6) and the absolute bifurcation limit (Eq. 3.9) for a modern digital AI accelerator, we parameterize a 1 mm² active tensor core block. The parameters reflect state-of-the-art projections for sub-2nm Complementary FET (CFET) nodes utilizing Backside Power Delivery Networks (BSPDN) as of 2026.

Table S1. Empirical Parameterization for a 1 mm² Sub-2nm CFET Tensor Core

Parameter	Symbol	Value	Physical Justification & Source
Intrinsic Time Constant	τ_0	$10^{-11}\ s$	Dominated by interconnect RC delays rather than intrinsic gate delays in sub-2nm nodes (Gupta et al., 2021)
Logic Switching Energy	E_{logic}	$2 \times 10^{-8}\ J/mm^2$	Energy required to flip physical states in CFET logic gates (Virmani & Chatterjee, 2025)
Supply Voltage	V_{DD}	0.7 V	Projected nominal operating voltage for sub-2nm CFET nodes (IRDS, 2023)
Total Switched Capacitance Density	$C_{int,total}$	$3.26 \times 10^{-7}\ F/mm^2$	Aggregated parasitic capacitance of the memory hierarchy and NoC routed within the 1 mm² block, assuming standard activity factors (Guo, et al. 2025)
Static Maintenance Power	P_{maint}	$5\ W/mm^2$	Subthreshold leakage evaluated at the effective junction temperature $T_{eff} \approx 345.5\ K$ (Roy et al., 2003; Gupta et al., 2021)
Thermal Conduction Loss	P_{cond}	$1\ W/mm^2$	Macroscopic thermal conduction through the packaging materials
Cross-Coupling Drag	P_{cross}	$2.5\ W/mm^2$	Thermoelectric leakage driven by Seebeck and Soret effects under extreme thermal gradients in dense 3D-ICs (Li et al., 2024)
Effective Thermal Resistance	$R_{th_{eff}}$	$0.8\ K \cdot mm^2/W$	Median empirical value for BSPDN architectures, which require extreme wafer thinning and lack bulk silicon heat spreaders (Xie & Wei, 2024).
Ambient Temperature	T_0	$293\ K$	Standard data center facility cooling baseline
Effective Junction Temp	T_{eff}	$345.5\ K$	Entropy-flux-preserving volumetric integral for the active area, preventing extreme but negligible hotspots from skewing the macroscopic exergy balance (Zhu et al., 2025)

Spatial data movement across the memory hierarchy dominates the dynamic energy budget, with the physical charging of parasitic interconnect capacitance yielding $E_{transport} = 1/2\, C_{int,total} {V_{DD}}^2 \approx 8 \times 10^{-8}\ J/mm^2$ (Guo, et al. 2025).

S4.2. Formulation of the Iterative Master Equation

Because the exact optimality condition is a transcendental equation (with the operation time τ present implicitly within the Thermal Brake penalty $P_{penalty}$, it cannot be solved via simple arithmetic. We must formulate it as an iterative dynamical system.

Baseline Power Density (Eq. 15):

$$P_{base} = P_{maint} + P_{cond} + P_{cross} = 8.5\ W/mm^2$$

First, we define the combined dynamic switching parameter:

$$\left(E_{logic} + E_{transport}\right)\tau_0 = 10^{-18}\ J \cdot s/mm^2$$

To simplify the numerical evaluation, we express the operation time τ in nanoseconds, defining t such that $\tau = t \cdot 10^{-9}\ s$. Because the combined switching parameter is exactly $10^{-18}\ J \cdot s/mm^2$, the dynamic power density as a function of t elegantly simplifies due to the cancellation of the 10^{-18} scaling factors:

$$P_{dyn} = \frac{\left(E_{logic} + E_{transport}\right)\tau_0}{\tau^2} = \frac{10^{-18}}{(t \cdot 10^{-9})^2} = \frac{1}{t^2}\ W/mm^2$$

Next, we expand the components of the Thermal Brake penalty

$$P_{penalty} = |\alpha|\left(\delta + \frac{A_{total}}{T_{eff}}\right) \qquad (S4.1)$$

Transient Thermal Impedance ($|\alpha|$):

$$|\alpha(t)| = \left|\frac{\partial P_{dyn}}{\partial \tau}\right| R_{th_{eff}} = \frac{1.6 \times 10^9}{t^3}\ K/s$$

Thermal-Leakage Feedback (δ): In sub-2nm nodes, subthreshold leakage doubles approximately every 12 K. The temperature derivative is

$$\frac{\partial \mathrm{P}_{maint}}{\partial T_{eff}} \approx 0.289\ W/(K \cdot mm^2)$$

$$\delta(t) = \tau \frac{\partial \mathrm{P}_{maint}}{\partial T_{eff}} = 0.289 \times 10^{-9} t\ J/(K \cdot mm^2)$$

Total Anergy per Operation (A_{total}):

$$A_{total}(t) = (E_{dyn} + E_{base})\tau = \left(\frac{1}{t} + 8.5t\right) \times 10^{-9}\ J/mm^2$$

Substituting these into the penalty equation and dividing by $T_{eff} = 345.5\ K$ yields a compact polynomial function for the Thermal Brake:

$$P_{penalty}(t) = \frac{0.50176}{t^2} + \frac{0.00463}{t^4}\ W/mm^2$$

The iterative sequence to find the true optimal operation time t_{n+1} is therefore:

$$t_{n+1} = \sqrt{\frac{1}{8.5 - P_{penalty}(t)}} \qquad (S4.2)$$

S4.3. 10-Step Numerical Convergence (The Damped Oscillator)

We initialize the system at **Iteration 0** using the naive isothermal optimum, assuming no thermal brake ($P_{penalty} = 0$). $t_{iso} \approx 0.343\ ns$ (corresponding to $f_{iso} \approx 2.92\ GHz$).

Table S2. Iterative Convergence of the Thermodynamic Optimum

Iteration (n)	**Input Time t_n (ns)**	**Thermal Brake $P_{penalty}\ W/mm^2$**	**Effective Dissipation $D_{exact}\ W/mm^2$**	**Output Time t_{n+1} (ns)**	**Clock Frequency f_n (GHz)**
0 (Naive)	**0.343**	4.600	3.900	**0.506**	2.92→1.98
1	**0.506**	2.030	6.470	**0.393**	1.98→2.54
2	**0.393**	3.442	5.058	**0.445**	2.54→2.25
3	**0.445**	2.652	5.848	**0.414**	2.25→2.42
4	**0.414**	3.085	5.415	**0.430**	2.42→2.33
5	**0.430**	2.849	5.651	**0.421**	2.33→2.38
6	**0.421**	2.978	5.522	**0.426**	2.38→2.35
7	**0.426**	2.906	5.594	**0.423**	2.35→2.36
8	**0.423**	2.949	5.551	**0.424**	2.36
9	**0.424**	2.935	5.565	**0.424**	2.36
10 (Converged)	**0.424**	**2.935**	**5.565**	**0.424**	**2.36**

Physical Interpretation of the Convergence:

This numerical evaluation perfectly captures the non-linear dynamical feedback loop (Le Chatelier's principle) governing modern processors.

1. When the system attempts to run at the naive 2.92 GHz (t_0), the massive transient heating triggers a severe thermal penalty of 4.6 W/mm^2. The system 'hits the brakes,' collapsing the effective dissipation and forcing the frequency down to 1.98 GHz.

2. At 1.98 GHz (t_1), the dynamic power drops significantly. The transient thermal impedance relaxes, reducing the thermal penalty to 2.03 W/mm^2. The system 'sees' available thermal headroom and accelerates to 2.54 GHz.
3. This oscillatory 'throttling' dampens over successive iterations, ultimately converging to a stable thermodynamic attractor at $\tau_{opt}^{*} \approx 0.424\ ns$ $(2.36\ GHz)$. At this equilibrium, the Thermal Brake continuously consumes 34.5% of the package's effective dissipation capacity.

S4.4. Calculation of the Absolute Bifurcation Limit (Thermal Runaway)

To find the absolute physical speed limit where the system undergoes a saddle-node bifurcation ($\mathcal{M} \to 0$), we evaluate the residual thermal capacity of the packaging (C_{res}) using Equation 3.7:

$$C_{res} \equiv \frac{T_0}{R_{th_{eff}}} - \mathrm{P}_{base} - T_{eff}\frac{\partial \mathrm{P}_{base}}{\partial T_{eff}} \approx 257.9\ W/mm^2$$

Using Equation 3.9, we calculate the absolute thermal runaway limit. Note that the squared informational anergy term (${\mathrm{A}_{info}}^2 \approx 7.84 \times 10^{-28}$) is negligibly small compared to the dynamic transport term under the square root. Thus, the square root term evaluates to $\sqrt{4 \times 257.9 \times 10^{-18}} \approx 3.2118 \times 10^{-8}\ J/mm^2$. Substituting these values yields:

$$\tau_{runaway} = \frac{2.8 \times 10^{-14} + 3.2118 \times 10^{-8}}{2 \times 257.9} \approx 6.226 \times 10^{-11}\ s = 0.06226\ ns.$$

$$f_{runaway} \approx 16.1\ GHz$$

This establishes an absolute, uncrossable physical wall. Attempting to drive the clock beyond 16.1 GHz results in a divergent temperature increase, as the dynamic power density exceeds the residual thermal capacity of the package, destroying the steady-state thermal equilibrium.

S4.5. Calculation of Macroscopic Exergetic Losses at Frequency Limits

To evaluate the total macroscopic exergy destroyed per operation (Ex_{opt}), we use the generalized penalty function derived in Eq. S4.3 for the direct dissipation regime:

$$\Lambda(T_{eff}, \chi) = \frac{T_{eff}}{T_0}\left[1 + \chi\left(1 - \frac{T_0}{T_{eff}}\right)\right] \qquad (S4.3)$$

Assuming an ambient temperature $T_0 = 293\ K$ and a facility cooling inefficiency factor $\chi = 0.1$.

1. Evaluation at the Thermodynamic Optimum ($f_{opt}^{*} \approx 2.36\ GHz$, $\tau_{opt}^{*} \approx 0.424\ ns$):

Dynamic Power: $P_{dyn} = \left(E_{logic} + E_{transport}\right)\tau_0/\tau^2 \approx 5.56\ W/mm^2$

Total Power: $P_{total} = P_{dyn} + P_{base} \approx 14.1\, W/mm^2$

Energy per Operation: $E_{op} = P_{total}\tau^*_{opt} \approx 5.96 \times 10^{-9}\ J/mm^2$

Effective Temperature: $T_{eff} \approx 345.5\, K$

Penalty Multiplier: $\Lambda \approx 1.197$

Total Exergy Cost: $Ex_{op}(f^*_{opt}) = E_{op}\Lambda \approx 7.13 \times 10^{-9}\ J/mm^2$

2. Evaluation at the Saddle-Node Bifurcation Limit ($f_{runaway} \approx 16.1\, GHz$, $\tau_{runaway} \approx 0.0623\, ns$):

Dynamic Power: $P_{dyn} = (E_{logic} + E_{transport})\,\tau_0/\tau^2 \approx 257.6\, W/mm^2$

Total Power at Bifurcation: At the runaway threshold, the total power approaches the maximum cooling capacity of the package: $P_{total} \approx T_0/R_{th_{eff}} \approx 366.25\, W/mm^2$

Energy per Operation: $E_{op} = P_{total}\tau \approx 22.8 \times 10^{-9}\ J/mm^2$

Effective Temperature at Bifurcation: $T_{runaway} \approx T_0 + P_{total}R_{th_{eff}} \approx 586\, K$. (Note: In physical reality, the chip would melt before reaching this steady-state, but this represents the mathematical asymptote).

Penalty Multiplier: $\Lambda \approx 2.10$

Total Exergy Cost: $Ex_{op}(f_{runaway}) \approx 47.88\ nJ/mm^2$

Comparison: The ratio of exergy destroyed is 47.88/7.13≈6.7. This rigorous calculation demonstrates that operating at the runaway limit requires nearly an order of magnitude more macroscopic work per logical operation than operating at the thermodynamic optimum, validating the severe exergetic penalty of the Speed-Efficiency Paradox.

S4.6. Empirical Parameterization and Bifurcation Analysis for Analog ReRAM Crossbars

To rigorously evaluate the exact optimality condition and the absolute bifurcation limit for analog in-memory computing (CiM) discussed in Section 4.2 of the main text, we parameterize a dense memristive crossbar array (e.g., a state-of-the-art 2026 TaOx/HfOx ReRAM array). Unlike digital systems where the energy is dominated by spatial transport, analog systems face a fundamental trade-off between the continuous-state precision (N) and the localized Joule heating required to maintain the conductive filament against thermal noise.

The parameters below reflect state-of-the-art projections for high-density 3D passive crossbar arrays.

Table S3. Empirical Parameterization for a Dense ReRAM Crossbar Array

Parameter	Symbol	Value	Physical Justification & Source
Cell Footprint Area	A_{cell}	$10^{-7}\ mm^2$	Extremely high spatial density in modern 3D crossbars, severely restricting lateral heat spreading (Schmid et al., 2024; Kadyrov et al., 2026)
Parasitic Wire Capacitance	C_{wire}	200 aF	Wordline and bitline fringe capacitances in dense passive arrays (Pande et al., 2024)
Programming Voltage	V_{prog}	2 V	Typical voltage required for SET/RESET pulses in TaOx/HfOx heterostructures (Ielmini & Pedretti, 2025); comparable ultra-low-voltage operation has been demonstrated in ZrOx/TaOx bilayers (Lee et al., 2026)
Substrate Confinement Constant	c	$5.08 \times 10^{-21}\ J/K$	Derived from the filament -dissipation theorem, where $c = \eta k_B$. The dimensionless structural factor $\eta \approx 368$ accounts for the filament formation energy barriers and activation volume in TaOx/HfOx heterostructures formation energy barriers and material-specific activation energy (Xia & Yang, 2019; Jiang et al., 2022)
Effective Thermal Resistance	$R_{th_{eff}}$	$1.0\ K \cdot mm^2/W$	Extracted from electro-thermal models of dense passive arrays lacking active bulk cooling (Oprins et al., 2022; Virmani & Chatterjee, 2025)
Maximum Operating Temp.	T_{max}	400 K	The physical boundary before irreversible thermal degradation of the TaOx/HfOx filament occurs (Virmani & Chatterjee, 2025)
Ambient Baseline Temp.	T_0	300 K	Standard room-temperature baseline for the crossbar substrate.

Formulation of the Dual-Bottleneck Master Equation

In a physical analog system, the total dynamic energy density $E_{dyn,total}$ dissipated into the substrate per operation is the sum of the precision-dependent structural anergy (which constitutes the logical switching energy, E_{logic}) and the precision-independent parasitic wire charging (which constitutes the spatial transport energy, $E_{transport}$):

$$E_{dyn,total} = E_{logic} + E_{transport} = cN^2T_{eff} + \frac{1}{2}C_{wire}{V_{prog}}^2 \qquad (S4.4)$$

where c is the material-specific confinement constant derived from the filament formation energy barriers (Xia & Yang, 2019; Jiang et al., 2022). Using the parameters from Table S3, the constant spatial transport energy per cell is $E_{transport} = 0.4\ fJ$. The parasitic thermal baseline is therefore $E_{transport}R_{th_{eff}}/A_{cell} = 4\ ns/K$.

When the crossbar operates at a finite time τ, the localized power dissipation induces a transient temperature spike. The effective junction temperature T_{eff} becomes dynamically coupled to the operation speed, creating a positive feedback loop:

$$T_{eff} = T_0 + \frac{cN^2T_{eff} + E_{transport}}{\tau A_{cell}}R_{th_{eff}} \qquad (S4.5)$$

Solving for T_{eff} yields the exact thermal runaway function for analog precision:

$$T_{eff} = \frac{T_0 + \frac{E_{transport} R_{th_{eff}}}{\tau A_{cell}}}{1 - \frac{\tau_{runaway}}{\tau}} \qquad (S4.6)$$

where $\tau_{runaway} = cN^2 R_{th_{eff}} / A_{cell}$ is the absolute thermodynamic speed limit. Just as in the digital domain (Section 3.2), this represents a saddle-node bifurcation where the required logical energy diverges, destroying the steady-state thermal equilibrium.

While $\tau_{runaway}$ defines the absolute mathematical asymptote, physical operation is strictly bounded by the material degradation threshold $T_{max} \approx 400\ K$. Solving Eq. S4.6 for $T_{eff} = T_{max}$ yields the material-constrained operation time, $\tau_{T_{max}}$:

$$\tau_{T_{max}} = \frac{\tau_{runaway} T_{max} + \frac{E_{transport} R_{th_{eff}}}{A_{cell}}}{T_{max} - T_0} \qquad (S4.7)$$

Step-by-Step Numerical Evaluation: The Precision-Temperature Trap

To demonstrate the catastrophic impact of the Precision-Temperature Trap, we evaluate Eq. S4.4 for two distinct precision regimes.

Scenario A: High-Precision (8-bit, $N = 256$)

Structural Energy Dominance: The required filament stiffness scales quadratically. $\tau_{runaway} \approx 3.33\ ns$.

Absolute Runaway Frequency: $f_{runaway} = 1/\tau_{runaway} \approx 300\ MHz$

Material Limit: $\tau_{T_{max}} \approx 13.36\ ns$, restricting practical operation to $f_{max} = 1/\tau_{T_{max}} \approx 75\ MHz$.

Scenario B: Low-Precision (4-bit, $N = 16$)

Parasitic Energy Dominance: The structural programming energy drops significantly ($E_{logic} \approx 0.39\ fJ$), becoming comparable to the precision-independent $E_{transport} \approx 0.4\ fJ$. $\tau_{runaway} \approx 13\ ps$.

Absolute Runaway Frequency: $f_{runaway} \approx 76.9\ GHz$

Material Limit: $\tau_{T_{max}} \approx 92\ ps$, restricting practical operation to $f_{max} \approx 10.8\ GHz$.

Table S4. Comparative Thermodynamic Limits of Analog Precision

Precision	4-bit	8-bit
States (N)	16	256
Logical Energy (E_{logic})	~0.39 fJ	~100 fJ
Spatial Transport Energy ($E_{transport}$)	0.4 fJ	0.4 fJ
Bifurcation Limit ($\tau_{runaway}$)	13 ps	3.33 ns
Absolute Runaway Frequency ($f_{runaway}$)	76.9 GHz	300 MHz
Material Limit at T_{max} ($\tau_{T_{max}}$)	92 ps	13.36 ns
Max Clock Frequency (f_{max})	10.8 GHz	75 MHz

Physical Interpretation:

Table S4 rigorously demonstrates that the 'Precision-Temperature Trap' is a Dual-Bottleneck. At low precisions (4-bit), the system is bottlenecked by the precision-independent thermal floor imposed by spatial transport capacitance ($E_{transport}$). Here, the analog substrate is highly efficient, boasting a theoretical runaway frequency ($f_{runaway} \approx 76{,}9\ GHz$) well above digital CFETs.

However, attempting to scale to digital-like precision (8-bit) triggers an exponential explosion in logical anergy (E_{logic}), shifting the bottleneck entirely to localized Joule heating. This forces the absolute thermodynamic bifurcation limit to collapse to ~300 MHz. Furthermore, because physical devices suffer irreversible filament degradation long before reaching this mathematical asymptote, the system is strictly constrained by the Material Limit (T_{max}). To prevent thermal damage, the maximum practical frequency ($f_{max} = 1/\tau_{T_{max}}$) must be throttled down from ~10.8 GHz (at 4-bit) to a mere ~75 MHz (at 8-bit). This rigorously proves that high-precision analog arrays cannot sustain high-throughput operations without triggering catastrophic thermal runaway or material failure.

References

Aghaee Rad, H., et al. (2025). 'Scaling and networking a modular photonic quantum computer.' *Nature*, 638, 912–919. DOI: 10.1038/s41586-024-08406-9

Bejan, A. (1996). *Entropy Generation Minimization : The Method of Thermodynamic Optimization on Finite-Size Systems and Finite-Time Processes.* CRC Press, Boca Raton.

Berut, A., et al. (2012). 'Experimental verification of the Landauer principle.' *Nature*, 483(7388), 187-189. DOI: 10.1038/nature10872

Berx, J., Olander, T., & Kirchberg, H. (2026). 'Pareto-optimal work extraction and the thermodynamic cost of precision in quantum information engines.' *arXiv preprint* arXiv:2607.14973v1. DOI: 10.48550/arXiv.2607.14973

Callen, H. B. (1985). *Thermodynamics and an Introduction to Thermostatistics* (2nd ed.). John Wiley & Sons.

Chaudhry, A., Kumar, M.J. (2004). 'Controlling short-channel effects in deep-submicron SOI MOSFETs for improved reliability: a review.' *IEEE Transactions on Device and Materials Reliability*. 4(1), 99-109. DOI: 10.1109/TDMR.2004.824359

Chattopadhyay, P., Misra, A., Pandit, T., & Paul, G. (2025). 'Landauer Principle and Thermodynamics of Computation'. *Reports on Progress in Physics*, 88(8), 086001. DOI: 10.1088/1361-6633/add6b3.

Chen, R., & Craven, G. T. (2025). 'Neuromorphic heat transport effects in a molecular junction.' *The Journal of Chemical Physics*, 163(5), 054115. DOI: 10.1063/5.0274613

Chen, S., et al. (2024). *'Thermal Implications in Scaling High-Performance Server 3D Chiplet-Based 2.5D SoC from FinFET to Nanosheet.*' 2024 IEEE Computer Society Annual Symposium on VLSI (ISVLSI). DOI:10.1109/ISVLSI61997.2024.00020

Cheng, W., et al. (2024). 'Thermodynamic analysis and optimization of a condensation-driven dilution refrigerator.' *Journal of Applied Physics*. 135(23), 235901. DOI: 10.1063/5.0203218.

Chimal-Eguia, J.C., et al. (2025). 'Thermodynamics of a Simple Endoreversible Model for Computer Gates.' *Mathematics*, 13(16), 2577. DOI: 10.3390/math13162577

Cobos-Murcia, J. (2026). 'Formulation of the Principle of Energy Quality: Toward a Fourth Law of Thermodynamics.' *Chemical Thermodynamics and Thermal Analysis*. 21. 100265. DOI: 10.1016/j.ctta.2026.100265

Curzon, F.L., Ahlborn, B. (1975). 'Efficiency of a Carnot engine at maximum power output.' *American Journal of Physics*, 43(1), 22–24. DOI: 10.1119/1.10023.

Das, S., et al. (2024). 'Bandwidth-Latency-Thermal Co-Optimization of Interconnect-Dominated Many-Core 3D-IC.' *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 33(2), 346-357. DOI: 10.1109/TVLSI.2024.3467148.

DeBenedictis, E.P. (2025). 'Scaling up Reversible Logic with HKI Superconducting Inductors.'. *arXiv preprint*, DOI: 10.48550/arXiv.2504.09229v2.

DeBenedictis, E.P., Mee, J.K., Frank, M.P. (2017). 'The Opportunities and Controversies of Reversible Computing.' *Computer*. 50(6). 76-80. DOI: 10.1109/MC.2017.177.

Diamantini, M. C., Gammaitoni, L., & Trugenberger, C. A. (2016). 'Landauer bound for analog computing systems.' *Physical Review E*, 94(1), 012139. DOI: 10.1103/PhysRevE.94.012139

Dincer, I., & Rosen, M.A. (2020). *Exergy: Energy, Environment and Sustainable Development (3rd ed.).* Elsevier Science. 724p. ISBN: 9780128243725.

Dutta, D., et al. (2025). 'Deep-Cryogenic Modeling of 22-nm FDSOI MOSFETs based on BSIM-IMG.' *arXiv preprint*, arXiv:2511.22297v1.

Ebrahimi, K., et al. (2014). 'A review of data center cooling technology, operating conditions and the corresponding low-grade waste heat recovery opportunities.' *Renewable and Sustainable Energy Reviews*, 31. 622-638. DOI: 10.1016/j.rser.2013.12.007

Esmaeilzadeh, H., et al. (2011). 'Dark silicon and the end of multicore scaling.' *ACM Sigarch Computer Architecture News.* 39(3):365-376. DOI: 10.1145/2024723.2000108

Falasco, G., & Esposito, M. (2025). 'Macroscopic stochastic thermodynamics.'

Reviews of Modern Physics, 97(1), 015002. DOI: 10.1103/RevModPhys.97.015002

Frank, M.P. (2017). 'Foundations of generalized reversible computing.' In *International Conference on Reversible Computation*, pp. 19-34. Springer International Publishing.

van Gerven, M. (2026). 'Dynamical systems foundations for neuromorphic intelligence.' *Neuromorphic Computing and Engineering*. 6. DOI: 10.1088/2634-4386/ae5380

Gouy, G. (1889) 'Sur l'énergie utilisable.' *Journal de Physique*, 8, 501-518. DOI: 10.1051/jphystap:018890080050101

Guha Majumdar, M. (2026). 'Thermodynamic Limits of Fault-Tolerant Quantum Computing Beyond the Weak-Coupling, Quasistatic Regime.' *Entropy*, 28(5), 546. DOI: 10.3390/e28050546

Guo, F. et al. (2025) 'Recent Developments in Thermal Management of 3D ICs: A Review.' in *IEEE Access*, vol. 13, pp. 94286-94301, 2025, DOI: 10.1109/ACCESS.2025.3569879

Gupta, M. K., et al. (2021). 'The Complementary FET (CFET) 6T-SRAM.' *IEEE Transactions on Electron Devices*. vol. 68(12), 6106-6111. DOI:10.1109/TED.2021.3121349

Heydari A., et al. (2024). 'Experimental evaluation of direct-to-chip cold plate liquid cooling for high-heat-density data centers.' *Applied Thermal Engineering*, 239. 122122, DOI: 10.1016/j.applthermaleng.2023.122122

Holmes, D.S., Ripple, A.L., & Manheimer, M.A. (2013). 'Energy-efficient superconducting computing—Power budgets and requirements.' *IEEE Transactions on Applied Superconductivity*, 23(3), 1701610-1701610. DOI: 10.1109/TASC.2013.2244634.

IEA (2026), *Energy and AI in East Asia*, IEA, Paris https://www.iea.org/reports/energy-and-ai-in-east-asia, Licence: CC BY 4.0

Ielmini, D. & Pedretti, G. (2025). 'Resistive Switching Random-Access Memory (RRAM): Applications and Requirements for Memory and Computing.' *Chemical Reviews*. 125(12). 5584-5625. DOI: 10.1021/acs.chemrev.4c00845.

Ielmini, D., & Wong, H. S. P. (2018). 'In-memory computing with resistive switching devices.' *Nature Electronics*, 1(6), 333-343. DOI: 10.1038/s41928-018-0092-2

IRDS (2023). *International Roadmap for Devices and Systems: More Moore*. IEEE.

Jiang, J., et al. (2022). 'Filament Formation in TaOx Thin Films for Memristor Device Application: Modeling Electron Energy Loss Spectra and Electron Transport.' *Advanced Electronic Materials*, 9(1), 2200828. DOI: 10.1002/aelm.202200828

Jou, D., & Lebon, G. (2010). *Extended Irreversible Thermodynamics* (4th ed.). Springer Dordrecht. DOI: 10.1007/978-90-481-3074-0

Kadyrov, A., et al. (2026). 'Surface-enhanced thermal dissipation in 3D vertical resistive memory arrays with top selector transistors.' *Nanoscale Horizons*, 11(3):909-915. DOI: 10.1039/d5nh00653h

Kanbur, B.B. et al. (2021) 'System-level experimental investigations of the direct immersion cooling data center units with thermodynamic and thermoeconomic assessments,' *Energy, Elsevier*, vol. 217(C). DOI: 10.1016/j.energy.2020.119373.

Kim, K., et al. (2026) 'Governing Thermal Transport in Three-Dimensional Electronics.' *ACS Nano*. 20(24), 17119-17142. DOI: 10.1021/acsnano.6c01518.

Kim, S.K., & Kwon, K.-W. (2026). 'Electro-thermal analysis of CFETs with a dual-substrate architecture.' *Semiconductor Science and Technology*. 41(5). DOI:10.1088/1361-6641/ae6362.

Kim, S.J., & Yang, J.J. (2024). 'The critical role of thermodynamics in memristors.' *Device*, 2(12), 100625. DOI:10.1016/j.device.2024.100625

Krinner, S., Storz, S., Kurpiers, P. et al. (2019). 'Engineering cryogenic setups for 100-qubit scale superconducting circuit systems.' *EPJ Quantum Technol*. 6, 2. DOI: 10.1140/epjqt/s40507-019-0072-0

Landauer, R. (1961). Irreversibility and Heat Generation in the Computing Process.' *IBM Journal of Research and Development*, 5(3), 183–191. DOI: 10.1147/rd.53.0183.

Lee, M.K., et al. (2026). 'Fast and energy-efficient resistive switching in ZrOx/TaOx bilayer structures under ultra-low voltage operation for next-generation memory applications.' *Adv Compos Hybrid Mater 9*, 143. DOI: 10.1007/s42114-026-01695-w

Li, C., et al. (2024). 'Leakage and Thermal Reliability Optimization of Stacked Nanosheet Field-Effect Transistors.' *Micromachines*, 15(4), 424. DOI: 10.3390/mi15040424

Liao, M.-H., Hsieh, C.-P., & Lee, C.-C. (2017). 'Systematic investigation of self-heating effect on CMOS logic transistors from 20 to 5 nm technology nodes by experimental thermoelectric measurements and finite element modeling.' *IEEE Transactions on Electron Devices*, 64(2): 646-648. DOI:10.1109/TED.2016.2642404

Lipka-Bartosik, P., & Perarnau-Llobet, M. (2025). 'Minimizing Dissipation via Interacting Environments: Quadratic Convergence to Landauer Bound.' *Physical Review Letters*, 135(17), 170404. DOI: 10.1103/5w1r-1nzs.

Liu, Z., Wei, H., Cui, H., Sun, K., & Sun, B. (2024). 'Analysis of GAAFET's transient heat transport process based on phonon hydrodynamic equations.' *Acta Physica Sinica*, 73(14), 144401. DOI: 10.7498/aps.73.20240491.

Manzano, G., et al. (2024). 'Thermodynamics of Computations with Absolute Irreversibility, Unidirectional Transitions, and Stochastic Computation Times.' *Physical Review X*, 14(2), 021026. DOI: 10.1103/PhysRevX.14.021026

McGovern, J. (2015). 'Thermodynamic Analysis of Closed Steady or Cyclic Systems.' *Entropy*, 17(10), 6712-6742.

Mehonic, A., Ielmini, D., Roy, K., et al. (2024). 'Roadmap to neuromorphic computing with emerging technologies.' *APL Materials*, 12(10), 109201. DOI: 10.1063/5.0179424

Moore, A. L., & Shi, L. (2014).'Emerging challenges and materials for thermal management of electronics.' *Materials Today*, 17(4), 163-174. DOI: 10.1016/j.mattod.2014.04.003

Moran, M. J., et al. (2014). *Fundamentals of Engineering Thermodynamics (8th ed.)*. John Wiley & Sons.

Myeong, I., Song, I., Kang, M. J., & Shin, H. (2020). 'Self-Heating and Electrothermal Properties of Advanced Sub-5-nm Node Nanoplate FET.' *IEEE Electron Device Letters*, 41(7), 977-980. DOI: 10.1109/LED.2020.2998460

Oprins, H., et al. (2022). 'Package level thermal analysis of backside power delivery network (BS-PDN) configurations,' *2022 21st IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (iTherm)*, 1-7. DOI: 10.1109/iTherm54085.2022.9899603.

Onsager, L. (1931). 'Reciprocal Relations in Irreversible Processes.' *Physical Review*, 37(4), 405–426. DOI: 10.1103/PhysRev.37.405.

Oviedo, F., et al. (2025). 'Energy Use of AI Inference, Efficiency Pathways, and Test-Time Scaling.' *arXiv preprint arXiv*:2509.20241v2. URL: https://arxiv.org/abs/2509.20241

Pande, S., Chakrabarti, B., & Chakravorty, A. (2024). 'Thermal Crosstalk Analysis in ReRAM Passive Crossbar Arrays.' *2024 37th International Conference on VLSI Design and 2024 23rd International Conference on Embedded Systems (VLSID), IEEE Xplore*, pp. 390-395. DOI: 10.1109/VLSID60093.2024.00071

Parrondo, M.R., Horowitz, J.M., Sagawa, T. (2015). 'Thermodynamics of information.' *Nature Physics*, 11(2), 131-139. DOI: 10.1038/nphys3230

Peliti, L., & Pigolotti, S. (2021). *Stochastic Thermodynamics: An Introduction*. Princeton University Press. ISBN: 9780691215525.

Pop, E. (2010). 'Energy dissipation and transport in nanoscale devices.' *Nano Res*. 3, 147-169. DOI: 10.1007/s12274-010-1019-z

Proesmans, K., Ehrich, J., & Bechhoefer, J. (2020). 'Finite-time Landauer principle.' *Physical Review Letters*, 125(10), 100602. DOI: 10.1103/PhysRevLett.125.100602.

Rant, Z. (1956). 'Exergie, ein neues Wort fur 'technische Arbeitsfähigkeit.' *Forschung auf dem Gebiete des Ingenieurwesens*, 22(1), 36-37.

Rathore, S., Jaisawal, R. K., Kondekar, P. N., & Bagga, N. (2023). 'Demonstration of a Nanosheet FET With High Thermal Conductivity Material as Buried Oxide: Mitigation of Self-Heating Effect.' *IEEE Transactions on Electron Devices*, 70(4), 1970-1976. DOI: 10.1109/TED.2023.3241884

Rolandi, A. (2026). 'Optimal finite-time thermodynamics of effective two-level systems.' *Physical Review E*, 113(4), 044133, DOI: 10.1103/hv2n-5dcs

Roy, K., Mukhopadhyay, S., & Mahmoodi-Meimand, H. (2003). 'Leakage current mechanisms and leakage reduction techniques in deep-submicrometer CMOS circuits.' *Proceedings of the IEEE*. 91(2), 305-327. DOI: 10.1109/JPROC.2002.808156

Sagawa, T. (2012). *Thermodynamics of information processing in small systems.* Springer Tokyo. ISBN 978-4-431-54167-7. DOI:10.1007/978-4-431-54168-4

Salamon, P., & Berry, R. S. (1983). 'Thermodynamic length and dissipated availability.' *Physical Review Letters*, 51(13), 1127. DOI: 10.1103/PhysRevLett.51.1127

Schmid, S. W., et al. (2024). 'Picosecond Femtojoule Resistive Switching in Nanoscale VO2

Memristors.' *ACS Nano*, 18(33), 21966-21974. DOI: 10.1021/acsnano.4c03840

Shah, A.J., et al. (2008). 'Exergy Analysis of Data Center Thermal Management Systems.' ASME. *J. Heat Transfer*. February 2008; 130(2): 021401. DOI: 10.1115/1.2787024.

Shastri, B., et al. (2020). 'Photonics for artificial intelligence and neuromorphic computing.' *Nature Photonics*, 15, 102–114. DOI: 10.1038/s41566-020-00754-y

Shooshtari, M., et al. (2025). 'Review of Memristors for In-Memory Computing and Spiking Neural Networks.' *Advanced Intelligent Systems*. 000:e202500806. DOI: 10.1002/aisy.202500806

Seifert, U. (2025). *Stochastic Thermodynamics*. Cambridge University Press. ISBN: 9781316519554.

Seshanth, R., et al. (2026). 'Mapping of the thermal profile and leakage current behavior in scaled Forksheet FET.' *Journal of Computational Electronics*. 25(2). DOI: 10.1007/s10825-026-02521-8

van Staveren, J., et al. (2025). *'Cryo-CMOS Bias-Voltage Generation and Demultiplexing at mK Temperatures for Large-Scale Arrays of Quantum Devices.*' IEEE Transactions on Quantum Engineering. DOI: 10.1109/TQE.2025.3580377.

Stodola, A. (1904). *Die Dampfturbinen mit einem Anhange über die Aussichten der Wärmekraftmaschinen und über die Gasturbine*. Springer Berlin, Heidelberg. DOI: 10.1007/978-3-642-92293-0.

Strogatz, S. H. (2015). *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering.* Westview Press.

Swoboda, T., et al. (2023). 'Spatially-Resolved Thermometry of Filamentary Nanoscale ReRAM Devices.' *ACS Applied Electronic Materials*, 5(9), 5025–5031. DOI: 10.1021/acsaelm.3c00782

Szilard, L. (1929). 'Über die Entropieverminderung in einem thermodynamischen System bei Eingriffen intelligenter Wesen.' *Zeitschrift für Physik*, 53(11-12), 840-856. DOI: 10.1007/BF01341281.

Tkachenko, A.V. (2025). 'Thermodynamic bounds on energy use in Deep Neural Networks.' *arXiv preprint*, arXiv:2503.09980.

Trupiano, G., et al. (2026). 'Thermally modulated SINIS transconductance amplifier.' *Physical Review Applied (APS)*. 25(3). 034087. DOI: 10.1103/cz4n-rh4r

Virmani, D., & Chatterjee, B. (2025). Thermal Management Challenges in 2.5D and 3D Chiplet Integration: A Review on Architecture–Cooling Co-Design. *Eng*, 6(12), 373. DOI: 10.3390/eng6120373

Wiśniewski, P., et al. (2025). 'Silicon-oxide resistive switching memory based on the HSQ layer.' *Solid-State Electronics*. 230, 109223, DOI: 10.1016/j.sse.2025.109223

Wolpert, D. H., et al. (2024). 'Is stochastic thermodynamics the key to understanding the energy costs of computation?'. *Proceedings of the National Academy of Sciences (PNAS)*, 121(45), e2409370121. DOI: 10.1073/pnas.2321112121

Wong, H.S.P., & Salahuddin, S. (2015). 'Memory leads the way to better computing.' *Nature Nanotechnology*, 10(3), 191–194. DOI:10.1038/nnano.2015.29

Xie, F., & Wei, T. (2024). 'Thermal Mitigation Strategy for Backside Power Delivery Network.' *2024 IEEE 74th Electronic Components and Technology Conference (ECTC)*. 1485-1492. DOI:10.1109/ECTC51529.2024.00241.

Xia, Q., & Yang, J. J. (2019). 'Memristive crossbar arrays for brain-inspired computing.' *Nature Materials*, 18(4), 309-323. DOI: 10.1038/s41563-019-0291-x

Yadav, A., Caravelli, F., & Wolpert, D. (2026). 'Entropy production bounds for systems running computer programs.' *PNAS Nexus*. 5(4):pgag116. DOI: 10.1093/pnasnexus/pgag116

Yusupov, A., Atamuratov, A., & Khalilloev, M. (2025). 'Self-Heating Effects in Nanoscale MOSFETs.' In book*: Field-Effect Transistors - Fundamentals, Technologies, and Future Innovations* [Working Title] Publisher: IntechOpen, Editors: Kenan Cicek. DOI: 10.5772/intechopen.1012444

Zhang, C., et al. (2023). 'Energy saving potential analysis of two-phase immersion cooling system with multi-mode condenser,' *Applied Thermal Engineering*, 219, Part C. 119614. DOI: 10.1016/j.applthermaleng.2022.119614

Zhang, Y. et al. (2024). 'Fast Thermodynamic Study on a Silicon Nanotransistor at Cryogenic Temperatures.' *Nano letters*. 24. DOI: 10.1021/acs.nanolett.4c01424.

Zhu, K., et al. (2025). '3D-ICE 4.0: Accurate and efficient thermal modeling for 2.5D/3D heterogeneous chiplet systems.' *arXiv*: 2512.05823v1. DOI: 10.48550/arXiv.2512.05823